Optimal Hessian/Jacobian-Free Nonconvex-PL Bilevel Optimization

Feihu Huang [1 2]

Abstract

Bilevel optimization is widely applied in many machine learning tasks such as hyper-parameter learning, meta learning and reinforcement learning. Although many algorithms recently have been developed to solve the bilevel optimization problems, they generally rely on the (strongly) convex lower-level problems. More recently, some methods have been proposed to solve the nonconvex-PL bilevel optimization problems, where their upper-level problems are possibly nonconvex, and their lower-level problems are also possibly nonconvex while satisfying Polyak-Łojasiewicz (PL) condition. However, these methods still have a high convergence complexity or a high computation complexity such as requiring compute expensive Hessian/Jacobian matrices and its inverses. In the paper, thus, we propose an efficient Hessian/Jacobian-free method (i.e., HJFBiO) with the optimal convergence complexity to solve the nonconvex-PL bilevel problems. Theoretically, under some mild conditions, we prove that our HJFBiO method obtains an optimal convergence rate of $O(\frac{1}{T})$, where T denotes the number of iterations, and has an optimal gradient complexity of $O(\epsilon^{-1})$ in finding an ϵ-stationary solution. We conduct some numerical experiments on the bilevel PL game and hyper-representation learning task to demonstrate efficiency of our proposed method.

1. Introduction

Bilevel optimization (Colson et al., 2007; Liu et al., 2021a), as an effective two-level hierarchical optimization paradigm, is widely applied in many machine learning tasks such as hyper-parameter learning (Franceschi et al., 2018), meta learning (Franceschi et al., 2018; Ji et al., 2021) and reinforcement learning (Hong et al., 2020; Chakraborty et al., 2023). In the paper, we consider a class of nonconvex bilevel optimization problems:

$$\begin{aligned} &\min_{x\in\mathbb{R}^d, y\in y^*(x)} f(x,y)+\phi(x), \qquad \text{(Upper-Level)} \\ &\text{s.t. } y^*(x)\equiv \arg\min_{y\in\mathbb{R}^p} g(x,y), \qquad \text{(Lower-Level)} \end{aligned} \tag{1}$$

where the upper-level function $f(x,y)$ with $y \in y^*(x)$ is possibly nonconvex, and $\phi(x)$ is a convex but possibly nonsmooth regularization such as $\phi(x) = 0$ when $x \in \mathcal{X} \subseteq \mathbb{R}^d$ with convex set $\mathcal{X}$ otherwise $\phi(x) = +\infty$, or $\phi(x) = \|x\|_1$. The lower-level function $g(x,y)$ is possibly nonconvex on any y and satisfies Polyak-Łojasiewicz (PL) condition (Polyak, 1963), which relaxes the strong convexity. The PL condition is widely used to some machine learning models such as the over-parameterized deep neural networks (Frei & Gu, 2021; Song et al., 2021). In fact, Problem (1) widely appears in many machine learning tasks such as meta learning (Huang, 2023b) and reinforcement learning (Chakraborty et al., 2023).

The inherent nested nature of bilevel optimization gives rise to several difficulties in effectively solving these bilevel problems. For example, compared with the standard single-level optimization (i.e., $g\big(x,y\big) = 0$ in Problem (1)), the main difficulty of bilevel optimization is that the minimization of the upper and lower-level objectives are intertwined via the minimizer $y^*(x) \in \arg\min_y g(x,y)$ of the lower-level problem. To deal with this difficulty, recently many bilevel optimization methods (Ghadimi & Wang, 2018; Hong et al., 2020; Ji et al., 2021; Huang et al., 2022; Chen et al., 2023b) have been proposed by imposing the strong convexity assumption on the Lower-Level (LL) problems. The LL strong convexity assumption ensures the uniqueness of LL minimizer (i.e., LL Singleton), which simplifies both the optimization process and theoretical analysis, e.g., hyper-gradient $\nabla F(x)$ of the upper-level objective $F(x) = f(x, y^*(x))$ has a simple closed-form:

$$\begin{aligned} \nabla F(x) = &\nabla_x f(x,y^*(x)) \\ &- \nabla^2_{xy} g(x,y^*(x)) \nabla^2_{yy} g(x,y^*(x))^{-1} \nabla_y f(x,y^*(x)). \end{aligned} \tag{2}$$

Based on the above form of hyper-gradient $\nabla F(x)$, some gradient-based methods (Chen et al., 2022; Dagréou et al.,

[1]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China [2]MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing, China. Correspondence to: Feihu Huang <huangfeihu2018@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Table 1: Comparison of **gradient (or iteration) complexity** between our method and the existing methods in solving bilevel problem (1) for finding an ϵ-stationary solution ($\|\nabla F(x)\|^2 \leq \epsilon$ or its equivalent variants, where $F(x) = f(x, y)$ with $y \in y^*(x)$). Here $g(x, \cdot)$ denotes function on the second variable y with fixing variable x. **SC** stands for strongly convex. **H.J.F.** stands for Hessian/Jacobian-Free. **L.H.** stands for Lipschitz Hessian condition. *The Prox-F^{2}BA and F^{2}BA methods rely on some strict conditions such as Lipschitz Hessian of function $f(x, y)$.* **Note that** the GALET (Xiao et al., 2023) method simultaneously uses the PL condition, its Assumption 2 (i.e., let $\sigma_g = \inf_{x,y}\{\sigma^+_{\min}(\nabla^2_{yy} g(x, y))\} > 0$ for all (x, y)) and its Assumption 1 (i.e., $\nabla^2_{yy} g(x, y)$ is Lipschitz continuous). Clearly, when Hessian matrix $\nabla^2_{yy} g(x, y)$ is singular, its Assumption 1 and Assumption 2 imply that the lower bound of the non-zero singular values σ_g is close to zero (i.e., $\sigma_g \to 0$), under this case, the convergence results of the GALET are **meaningless**, e.g., the constant $L_w = \frac{\ell_{f,1}}{\sigma_g} + \frac{\sqrt{2}\ell_{g,2}\ell_{f,0}}{\sigma_g^2} \to +\infty$ used in its Lemmas 6 and 9. Under the other case, the PL condition, Lipschitz continuous of Hessian and its Assumption 2 (the singular values of Hessian is bounded away from 0, i.e., $\sigma_g > 0$) imply that GALET assumes strongly convex (Detailed discussion in the Appendix B).

Algorithm	**Reference**	$g(x, \cdot)$	**L.H.** on $f(\cdot, \cdot)$	**Complexity**	**Loop(s)**	**H.J.F.**
BOME	(Liu et al., 2022)	PL / local-PL		$O(\epsilon^{-1.5})$ / $O(\epsilon^{-2})$	Double	$\surd$
V-PBGD	(Shen & Chen, 2023)	PL / local-PL		$O(\epsilon^{-1.5})$ / $O(\epsilon^{-1.5})$	Double	$\surd$
GALET	(Xiao et al., 2023)	SC / PL		$O(\epsilon^{-1})$ / **Meaningless**	Triple	
SLM	(Lu, 2023)	PL / local-PL		$O(\epsilon^{-3.5})$ / $O(\epsilon^{-3.5})$	Double	$\surd$
Prox-F^{2}BA	(Kwon et al., 2023)	Proximal-EB	$\surd$	$O(\epsilon^{-1.5})$ / $O(\epsilon^{-1.5})$	Double	$\surd$
F^{2}BA	(Chen et al., 2024)	PL / local-PL	$\surd$	$O(\epsilon^{-1})$ / $O(\epsilon^{-1})$	Double	$\surd$
MGBiO	(Huang, 2023b)	PL / local-PL		$O(\epsilon^{-1})$ / $O(\epsilon^{-1})$	Single	
AdaPAG	(Huang, 2023a)	PL / local-PL		$O(\epsilon^{-1})$ / $O(\epsilon^{-1})$	Single	
HJFBiO	Ours	PL / local-PL		$O(\epsilon^{-1})$ / $O(\epsilon^{-1})$	Single	$\surd$

2022) require an explicit extraction of second-order information of $g(x, y)$ with a major focus on efficiently estimating its Jacobian and inverse Hessian. Meanwhile, the other gradient-based methods (Li et al., 2022; Dagréou et al., 2022; Sow et al., 2022b; Yang et al., 2023b) avoid directly estimating its second-order computation and only use the first-order information of both upper and lower objectives.

Recently, to relax the LL strong convexity assumption, another line of research is dedicated to bilevel optimization with convex LL problems, which bring about several challenges such as the presence of multiple LL local optimal solutions (i.e., Non-Singleton). Under this case, there does not exist the above hyper-gradient form (2). To handle this concern, some effective methods (Sow et al., 2022a; Liu et al., 2023; Lu & Mei, 2023; Cao et al., 2023) recently have been developed. For example, (Sow et al., 2022a) developed the primal-dual algorithms for bilevel optimization with multiple inner minima in the LL problem. Subsequently, (Lu & Mei, 2023) studied the constrained bilevel optimization with convex lower-level. Meanwhile, (Liu et al., 2023) proposed an effective averaged method of multipliers for bilevel optimization with convex lower-level.

In fact, the bilevel optimization problems with nonconvex LL problems frequently appear in many machine tasks such as hyper-parameter learning in training deep neural networks. Since the above bilevel optimization methods mainly rely on the restrictive LL strong convexity or convexity assumption, clearly, they can not effectively solve the bilevel optimization problems with nonconvex LL problems. Recently, some bilevel approaches (Liu et al., 2021b; 2022; Chen et al., 2023a; Liu et al., 2023; Huang, 2023b;a; Kwon et al., 2023; Chen et al., 2024) studied the bilevel optimization with non-convex lower-level. For example, (Liu et al., 2022) proposed an effective first-order method for nonconvex-PL bilevel optimization, where the lower-level problem is nonconex but satisfies PL condition. (Shen & Chen, 2023) designed an effective penalty-based gradient method for the constrained nonconvex-PL bilevel optimization. (Kwon et al., 2023) studied the nonconvex bilevel optimization with nonconvex lower-level satisfying proximal error-bound (EB) condition that is analogous to PL condition. Meanwhile, (Huang, 2023b) proposed a class of efficient momentum-based gradient methods for the nonconvex-PL bilevel optimization, which obtain an optimal gradient complexity but rely on requiring compute expensive projected Hessian/Jacobian matrices and its inverses. Subsequently, (Xiao et al., 2023) proposed a generalized alternating method (i.e., GALET) for nonconvex-PL bilevel optimization, which still relies on the expensive Hessian/Jacobian matrices. Unfortunately, the convergence results of the GALET method are meaningless (please see Table 1). Thus, there exists a natural question:

Could we propose an efficient Hessian/Jacobian-free method for the nonconvex-PL bilevel optimization with an optimal complexity?

In the paper, we affirmatively answer to this question, and propose an efficient Hessian/Jacobian-free method to solve Problem (1) based on the finite-difference estimator and a new projection operator. Our main contributions are given:

(i) We propose an efficient Hessian/Jacobian-free method (i.e., HJFBiO) based on the finite-difference estimator and a new useful projection operator. In particular, our HJFBiO method not only uses low computational first-order gradients instead of high computational Hessian/Jacobian matrices, but also applies the low computational new projection operator to vector variables instead of matrix variables. Thus, our HJFBiO method has a lower computation at each iteration.

(ii) We provide a solid convergence analysis for our HJFBiO method. Under some mild conditions, we prove that our HJFBiO method reaches the best known iteration (gradient) complexity of $O(\epsilon^{-1})$ for finding an ϵ-stationary solution of Problem (1), which matches the lower bound established by the first-order method for finding an ϵ-stationary point of nonconvex smooth optimization problems (Carmon et al., 2020).

(iii) We conduct some numerical experiments including bilevel Polyak-Łojasiewicz game and hyper-representation learning to demonstrate efficiency of our proposed method.

Meanwhile, (Chen et al., 2024) proposed a F^{2}BA method for the nonconvex-PL bilevel optimization, which also obtains the best known gradient complexity $O(\epsilon^{-1})$ for finding an ϵ-stationary solution of Problem (1), but it relies on some stricter conditions such as Lipschitz Hessian of the upper function $f(x,y)$. Under these strict conditions, although the F^{2}BA method (Chen et al., 2024) obtains a gradient complexity $O(\epsilon^{-1})$, this is not an optimal gradient complexity (Detailed discussion in the Appendix B).

Notations

Given function $f(x,y)$, $f(x,\cdot)$ denotes function *w.r.t.* the second variable with fixing x, and $f(\cdot,y)$ denotes function *w.r.t.* the first variable with fixing y. ∇_x denotes the partial derivative on variable x. Let $\nabla^2_{xy} = \nabla_x\nabla_y$ and $\nabla^2_{yy} = \nabla_y\nabla_y$. $\|\cdot\|$ denotes the ℓ_2 norm for vectors and spectral norm for matrices. $\langle x,y\rangle$ denotes the inner product of two vectors x and y. I_d denotes a d-dimensional identity matrix. $a_t = O(b_t)$ denotes that $a_t \leq cb_t$ for some constant $c>0$. The notation $\tilde{O}(\cdot)$ hides logarithmic terms.

$\mathcal{S}_{[\mu,L_g]}[\cdot]$ denotes a projection on the set $\{X\in\mathbb{R}^{d\times d}: \mu\leq \varrho(X)\leq L_g\}$, where $\varrho(\cdot)$ denotes the eigenvalue function. $\mathcal{S}_{[\mu,L_g]}$ can be implemented by using Singular Value Decomposition (SVD) and thresholding the singular values. $\mathcal{P}_{r_v}(\cdot)$ is a projection onto set $\{v\in\mathbb{R}^p: \|v\|\leq r_v>0\}$.

2. Preliminaries

In this section, we provide some mild assumptions and useful lemmas on the above Problem (1).

2.1. Mild Assumptions

Assumption 2.1. The function $g(x,\cdot)$ satisfies the Polyak-Łojasiewicz (PL) condition, if there exist $\mu>0$ such that for any given x, it holds that

$$\|\nabla_y g(x,y)\|^2 \geq 2\mu\big(g(x,y)-\min_y g(x,y)\big),\ \forall y\in\mathbb{R}^p.$$

Assumption 2.2. The function $g(x,y)$ is nonconvex and satisfies

$$\varrho\big(\nabla^2_{yy}g\big(x,y^*(x)\big)\big)\in[\mu,L_g], \tag{3}$$

where $y^*(x)\in\arg\min_y g(x,y)$, and $\varrho(\cdot)$ denotes the eigenvalue (or singular-value) function and $L_g\geq\mu>0$.

Assumption 2.3. The functions $f(x,y)$ and $g(x,y)$ satisfy

1) For all x,y, we have $\|\nabla_y f(x,y)\|\leq C_{fy}$, $\|\nabla^2_{xy}g(x,y)\|\leq C_{gxy}$;
2) The partial derivatives $\nabla_x f(x,y)$ and $\nabla_y f(x,y)$ are L_f-Lipschitz continuous;
3) The partial derivatives $\nabla_x g(x,y)$ and $\nabla_y g(x,y)$ are L_g-Lipschitz continuous.

Assumption 2.4. The partial derivatives $\nabla^2_{xy}g(x,y)$ and $\nabla^2_{yy}g(x,y)$ are L_{gxy}-Lipschitz and L_{gyy}-Lipschitz, e.g., for all $x,x_1,x_2\in\mathbb{R}^d$ and $y,y_1,y_2\in\mathbb{R}^p$

$$\|\nabla^2_{xy}g(x_1,y)-\nabla^2_{xy}g(x_2,y)\|\leq L_{gxy}\|x_1-x_2\|,$$
$$\|\nabla^2_{xy}g(x,y_1)-\nabla^2_{xy}g(x,y_2)\|\leq L_{gxy}\|y_1-y_2\|.$$

Assumption 2.5. The function $\Phi(x)=F(x)+\phi(x)$ is bounded below in $x\in\mathbb{R}^d$, *i.e.,* $\Phi^*=\inf_{x\in\mathbb{R}^d}\Phi(x)>-\infty$.

Assumption 2.1 is commonly used in bilevel optimization without the lower-level strongly convexity (Liu et al., 2022; Shen & Chen, 2023; Huang, 2023b). Assumption 2.2 ensures that the minimizer $y^*(x)=\arg\min_y g(x,y)$ is unique, which imposes the non-singularity of $\nabla^2_{yy}g(x,y)$ only at the minimizers $y^*(x)\in\arg\min_y g(x,y)$, as in (Huang, 2023b). *Based on Lemma G.6 of (Chen et al., 2024) given in the Appendix B, our Assumption 2.2 is reasonable when has an unique minimizer.* Meanwhile, we also study the case that $\min_y g(x,y)$ has multiple local minimizers in the following section 4.2. **Note that** since $y^*(x)\in\arg\min_y g(x,y)$, we can not have negative eigenvalues at the minimizer $y^*(x)$, so Assumption 2.2 assumes that $\varrho\big(\nabla^2_{yy}g\big(x,y^*(x)\big)\big)\in[\mu,L_g]$ instead of $\varrho\big(\nabla^2_{yy}g\big(x,y^*(x)\big)\big)\in[-L_g,-\mu]\cup[\mu,L_g]$. Since

$\nabla^2_{yy} g(x,y)$ is a symmetric matrix, its singular values are the absolute value of eigenvalues. Hence, we also can use $\varrho(\cdot)$ to denote the singular-value function.

Assumption 2.3 is commonly appeared in bilevel optimization methods (Ghadimi & Wang, 2018; Ji et al., 2021; Liu et al., 2022). Meanwhile, the BOME (Liu et al., 2022) uses the stricter assumption that $\|\nabla f(x,y)\|$, $\|\nabla g(x,y)\|$, $|f(x,y)|$ and $|g(x,y)|$ are bounded for any (x,y) in its Assumption 3. Assumption 2.4 is also commonly used in bilevel optimization methods (Ghadimi & Wang, 2018; Ji et al., 2021). Assumption 2.5 ensures the feasibility of the bilevel Problem (1).

For example, we consider a nonconvex-PL bilevel problem

$$\min_{x\in[1,2], y\in y^*(x)} \Big\{ f(x,y) = x^2 + y^2 + 3x\sin^2(y) \Big\}, \quad (4)$$
$$\text{s.t. } y^*(x) \equiv \min_{y\in\mathbb{R}} \Big\{ g(x,y) = xy^2 + x\sin^2(y) \Big\},$$

which can be rewritten as

$$\min_{x\in\mathbb{R}, y\in y^*(x)} \Big\{ f(x,y) + \phi(x) \Big\}, \quad (5)$$
$$\text{s.t. } y^*(x) \equiv \min_{y\in\mathbb{R}} \Big\{ g(x,y) = xy^2 + x\sin^2(y) \Big\},$$

where $\phi(x) = 0$ when $x \in [1,2]$ otherwise $\phi(x) = +\infty$. From the above bilevel problem (5), we can easily obtain $y^*(x) = 0$, $\nabla^2_{yy} g(x,y) = x\big(2+2\cos^2(y)-2\sin^2(y)\big)$, and then we have $\nabla^2_{yy} g(x,y^*(x)) = 4x > 0$ due to $x \in [1,2]$. Since $\nabla^2_{yy} g(x,y^*(x)) = 4x > 0$, our Assumption 2.2 holds. Meanwhile $\nabla^2_{yy} g(x,y) = x\big(2+2\cos^2(y)-2\sin^2(y)\big)$ for any $y \in \mathbb{R}$ may be zero or negative such as $\nabla^2_{yy} g(x,\pi/2) = 0$. Meanwhile, for any $y \in \mathbb{R}$, clearly $|f(x,y)|$ and $|g(x,y)|$ are not bounded. Thus, the assumptions used in (Liu et al., 2022; Xiao et al., 2023; Kwon et al., 2023) may be not satisfied.

2.2. Useful Lemmas

In this subsection, based on the above assumptions, we give some useful lemmas.

Lemma 2.6. *((Huang, 2023b)) Under the above Assumption 2.2, we have, for any $x \in \mathbb{R}^d$,*

$$\nabla F(x) = \nabla_x f(x, y^*(x)) - \nabla^2_{xy} g(x, y^*(x)) \Big[\nabla^2_{yy} g(x, y^*(x))\Big]^{-1} \nabla_y f(x, y^*(x)).$$

From the above Lemma 2.6, we can get the same form of hyper-gradient $\nabla F(x)$ as in (2). Since the Hessian matrix $\nabla^2_{yy} g(x,y)$ for all (x,y) may be singular, as in (Huang, 2023b), we define a useful hyper-gradient estimator:

$$\hat{\nabla} f(x,y) = \nabla_x f(x,y) - \nabla^2_{xy} g(x,y) \big(\mathcal{S}_{[\mu,L_g]}\big[\nabla^2_{yy} g(x,y)\big]\big)^{-1} \nabla_y f(x,y),$$

which replaces the standard hyper-gradient estimator $\breve{\nabla} f(x,y)$ used in (Ghadimi & Wang, 2018; Ji et al., 2021) for the strongly-convex lower-level optimization,

$$\breve{\nabla} f(x,y) = \nabla_x f(x,y) - \nabla^2_{xy} g(x,y) \big(\nabla^2_{yy} g(x,y)\big)^{-1} \nabla_y f(x,y).$$

Lemma 2.7. *((Huang, 2023b)) Under the above Assumptions 2.1-2.4, the functions (or mappings) $F(x) = f(x, y^*(x))$, $G(x) = g(x, y^*(x))$ and $y^*(x) \in \arg\min_{y\in\mathbb{R}^p} g(x,y)$ satisfy, for all $x_1, x_2 \in \mathbb{R}^d$,*

$$\begin{aligned}
&\|y^*(x_1) - y^*(x_2)\| \le \kappa \|x_1 - x_2\|,\\
&\|\nabla y^*(x_1) - \nabla y^*(x_2)\| \le L_y \|x_1 - x_2\|,\\
&\|\nabla F(x_1) - \nabla F(x_2)\| \le L_F \|x_1 - x_2\|,\\
&\|\nabla G(x_1) - \nabla G(x_2)\| \le L_G \|x_1 - x_2\|,
\end{aligned}$$

where $\kappa = C_{gxy}/\mu$, $L_y = \big(\frac{C_{gxy} L_{gyy}}{\mu^2} + \frac{L_{gxy}}{\mu}\big)\big(1 + \frac{C_{gxy}}{\mu}\big)$, $L_F = \Big(L_f + L_f\kappa + C_{fy}\big(\frac{C_{gxy} L_{gyy}}{\mu^2} + \frac{L_{gxy}}{\mu}\big)\Big)(1+\kappa)$ and $L_G = \Big(L_g + L_g\kappa + C_{gy}\big(\frac{C_{gxy} L_{gyy}}{\mu^2} + \frac{L_{gxy}}{\mu}\big)\Big)(1+\kappa)$.

Lemma 2.8. *((Huang, 2023b)) Let $\nabla F(x) = \nabla f(x, y^*(x))$ and $\hat{\nabla} f(x,y) = \nabla_x f(x,y) - \nabla^2_{xy} g(x,y) \big(\mathcal{S}_{[\mu,L_g]}\big[\nabla^2_{yy} g(x,y)\big]\big)^{-1} \nabla_y f(x,y)$, we have*

$$\|\hat{\nabla} f(x,y) - \nabla F(x)\|^2 \le \frac{2\hat{L}^2}{\mu} \big(g(x,y) - \min_y g(x,y)\big),$$

where $\hat{L}^2 = 4\big(L_f^2 + \frac{L_{gxy}^2 C_{fy}^2}{\mu^2} + \frac{L_{gyy}^2 C_{gxy}^2 C_{fy}^2}{\mu^4} + \frac{L_f^2 C_{gxy}^2}{\mu^2}\big)$.

3. Efficient Hessian/Jacobian-Free Bilevel Optimization Method

In the section, we propose an efficient Hessian/Jacobian-free method to solve the nonconvex-PL bilevel Problem (1) based on the finite-difference estimator and a new projection operator. Here we first define a useful projection operator:

Definition 3.1. Given matrix $H \in \mathbb{R}^{p\times p}$ and vector $v \in \mathbb{R}^p$, and $\mathcal{S}_{[\mu,L_g]}[\cdot]$ is a projection operator on the set $\{H \in \mathbb{R}^{p\times p} : \mu \le \varrho(H) \le L_g\}$ where $\varrho(\cdot)$ denotes the eigenvalue function, and $\mathcal{P}_{r_v}(\cdot)$ is a projection operator onto the set $\{v \in \mathbb{R}^p : \|v\| \le r_v\}$, then we define a **new projection operator** $\mathcal{M}_{r_h}(\cdot,\cdot)$ on set $\{H \in \mathbb{R}^{p\times p}, v \in \mathbb{R}^p : \|Hv\| \le r_h\}$, which satisfies

$$\mathcal{M}_{r_h}(H, v) \triangleq \mathcal{S}_{[\mu,L_g]}[H] \mathcal{P}_{r_v}(v), \quad (6)$$

where $0 < r_h \le r_v L_g$.

For notational simplicity, let $\mathcal{M}_{r_h}(H, v) = \mathcal{M}_{r_h}(Hv)$ in the following.

Algorithm 1 Hessian/Jacobian-free Bilevel Optimization (i.e, HJFBiO) Algorithm

1: **Input:** T, learning rates $\lambda > 0$, $\gamma > 0$, $\tau > 0$, and tuning parameters $\delta_\epsilon > 0$, $r_v > 0$, $r_h > 0$, and initial input $x_1 \in \mathbb{R}^d$, $y_1 \in \mathbb{R}^p$ and $v_1 \in \mathbb{R}^p$;
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Compute $u_t = \nabla_y g(x_t, y_t)$, and update $y_{t+1} = y_t - \lambda u_t$;
4: Compute $w_t = \widetilde{\nabla} f(x_t, y_t, v_t) = \nabla_x f(x_t, y_t) - \widetilde{J}(x_t, y_t, v_t, \delta_\epsilon)$, and update $x_{t+1} = \mathbb{P}^\gamma_{\phi(\cdot)}\big(x_t, w_t\big)$;
5: Compute $h_t = \widetilde{\nabla}_v R(x_t, y_t, v_t) = \mathcal{M}_{r_h}\big(\widetilde{H}(x_t, y_t, v_t, \delta_\epsilon)\big) - \nabla_y f(x_t, y_t)$, and update $v_{t+1} = \mathcal{P}_{r_v}\big(v_t - \tau h_t\big)$;
6: **end for**
7: **Output:** Chosen uniformly random from $\{x_t\}_{t=1}^T$.

From the above Lemma 2.6, the hyper-gradient $\nabla F(x)$ takes the form of

$$\nabla F(x) = \nabla_x f(x, y^*(x)) - \nabla^2_{xy} g(x, y^*(x))\big[\nabla^2_{yy} g(x, y^*(x))\big]^{-1} \nabla_y f(x, y^*(x)). \tag{7}$$

In the above problem (1), the lower objective function $g(x, y)$ on variable y is not strongly convex, so its Hessian matrix $\nabla^2_{yy} g(x, y)$ for all (x, y) may be singular. As in (Huang, 2023b), we define a useful hypergradient estimator:

$$\widehat{\nabla} f(x, y) = \nabla_x f(x, y) - \nabla^2_{xy} g(x, y)\big(\mathcal{S}_{[\mu, L_g]}\big[\nabla^2_{yy} g(x, y)\big]\big)^{-1} \nabla_y f(x, y). \tag{8}$$

Since the above hypergradient (8) requires computing the expensive projected Hessian inverse, as in (Yang et al., 2023b), we define a new hypergradient surrogates as follows:

$$\widehat{\nabla} f(x, y, v) = \nabla_x f(x, y) - \nabla^2_{xy} g(x, y) v, \tag{9}$$

where $v \in \mathbb{R}^p$ is an auxiliary vector to approximate the projected Hessian-inverse vector product $\big(\mathcal{S}_{[\mu, L_g]}\big[\nabla^2_{yy} g(x, y)\big]\big)^{-1} \nabla_y f(x, y)$ in (8), which can be rewritten as a solution of the following linear system:

$$v^* = \big(\mathcal{S}_{[\mu, L_g]}\big[\nabla^2_{yy} g(x, y)\big]\big)^{-1} \nabla_y f(x, y) = \arg\min_{v \in \mathbb{R}^p} \Big\{\frac{1}{2} v^T \mathcal{S}_{[\mu, L_g]}\big[\nabla^2_{yy} g(x, y)\big] v - v^T \nabla_y f(x, y)\Big\}. \tag{10}$$

Under this case, we can use the following new iterations to solve the nonconvex-PL bilevel problem (1): for $t \geq 1$,

$$\begin{cases} y_{t+1} = y_t - \lambda \nabla_y g(x_t, y_t), \\ x_{t+1} = \mathbb{P}^\gamma_{\phi(\cdot)}\big(x_t, \widehat{\nabla} f(x_t, y_t, v_t)\big), \\ v_{t+1} = \mathcal{P}_{r_v}\big(v_t - \tau \nabla_v R(x_t, y_t, v_t)\big), \end{cases} \tag{11}$$

where $\lambda > 0$, $\gamma > 0$ and $\tau > 0$ are learning rates, and the proximal operator defined as: given vectors $x_t, w_t \in \mathbb{R}^d$,

$$\mathbb{P}^\gamma_{\phi(\cdot)}\big(x_t, w_t\big) = \arg\min_{x \in \mathbb{R}^d} \big\{\langle w_t, x\rangle + \frac{1}{2\gamma}\|x - x_t\|^2 + \phi(x)\big\}. \tag{12}$$

In particular, on updating variable $v \in \mathbb{R}^p$, we use a projection $\mathcal{P}_{r_v}(\cdot)$ onto the set $\{v \in \mathbb{R}^p : \|v\| \leq r_v\}$ with $0 < r_v \leq \frac{C_{fy}}{\mu}$ to obtain the bounded variable v. Here we use the function

$$R(x, y, v) = \frac{1}{2} v^T \mathcal{S}_{[\mu, L_g]}\big[\nabla^2_{yy} g(x, y)\big] v - v^T \nabla_y f(x, y),$$

and then we have $\nabla_v R(x, y, v) = \mathcal{S}_{[\mu, L_g]}\big[\nabla^2_{yy} g(x, y)\big] v - \nabla_y f(x, y)$.

In the high-dimensional setting, clearly computing Hessian matrix $\nabla^2_{yy} g(x, y)$ and Jacobian matrix $\nabla^2_{xy} g(x, y)$ is expensive. To approximate this hypergradient efficiently, we further use the finite-difference technique to estimate the Hessian-vector $\nabla^2_{yy} g(x, y) v$ and Jacobian-vector $\nabla^2_{xy} g(x, y) v$ products. Specifically, given a small constant $\delta_\epsilon > 0$, we define two finite-difference estimators to estimate $\nabla^2_{yy} g(x, y) v$ and $\nabla^2_{xy} g(x, y) v$ respectively, defined as:

$$\widetilde{H}(x, y, v, \delta_\epsilon) = \frac{\nabla_y g(x, y + \delta_\epsilon v) - \nabla_y g(x, y - \delta_\epsilon v)}{2\delta_\epsilon}, \tag{13}$$

$$\widetilde{J}(x, y, v, \delta_\epsilon) = \frac{\nabla_x g(x, y + \delta_\epsilon v) - \nabla_x g(x, y - \delta_\epsilon v)}{2\delta_\epsilon}. \tag{14}$$

Then we can use the following Hessian/Jacobian-free iterations to solve the nonconvex-PL bilevel problem (1): at $(t+1)$-th iteration,

$$\begin{cases} y_{t+1} = y_t - \lambda \nabla_y g(x_t, y_t), \\ x_{t+1} = \mathbb{P}^\gamma_{\phi(\cdot)}\big(x_t, \widetilde{\nabla} f(x_t, y_t, v_t)\big), \\ v_{t+1} = \mathcal{P}_{r_v}\big(v_t - \tau \widetilde{\nabla}_v R(x_t, y_t, v_t)\big), \end{cases} \tag{15}$$

where $\lambda > 0$, $\gamma > 0$ and $\tau > 0$ are learning rates, and

$$\widetilde{\nabla} f(x_t, y_t, v_t) = \nabla_x f(x_t, y_t) - \widetilde{J}(x_t, y_t, v_t, \delta_\epsilon),$$
$$\widetilde{\nabla}_v R(x_t, y_t, v_t) = \mathcal{M}_{r_h}\big(\widetilde{H}(x_t, y_t, v_t, \delta_\epsilon)\big) - \nabla_y f(x_t, y_t).$$

Based on the above iterations (15), we give a procedure framework of our HJFBiO algorithm in Algorithm 1. When $\phi(x) \equiv 0$, in updating the variable x, we have $x_{t+1} = x_t - \gamma w_t = x_t - \gamma \widetilde{\nabla} f(x_t, y_t, v_t)$.

Note that in our Algorithm 1, we use two *low computational* finite-difference estimators (13) and (14) instead of computing *high computational* Hessian matrix $\nabla^2_{yy} g(x, y) \in$

$\mathbb{R}^{d\times d}$ and Jacobian matrix $\nabla^2_{xy}g(x,y)\in\mathbb{R}^{d\times p}$. Moreover, we also use a *low computational* projection operator $\{||\widetilde{H}(x,y,v,\delta_\epsilon)||\le r_h\}$ on vector $\widetilde{H}(x,y,v,\delta_\epsilon)$ instead of computing *high computational* projection operator $\mathcal{S}_{[\mu,L_g]}[\nabla^2_{yy}g(x,y)]$ used in (Huang, 2023b;a). Thus, our HJFBiO algorithm only requires a low computational complexity of $O(p+d)$ at each iteration.

From Algorithm 1, since $v_t=\mathcal{P}_{r_v}(v_t)$, we have

$$\begin{aligned}
&\mathcal{M}_{r_h}\big(\widetilde{H}(x_t,y_t,v_t,\delta_\epsilon)\big)\\
&=\mathcal{M}_{r_h}\Big(\frac{1}{2\delta_\epsilon}\big(\nabla_y g(x_t,y_t+\delta_\epsilon v_t)-\nabla_y g(x_t,y_t-\delta_\epsilon v_t)\big)\Big)\\
&=\mathcal{M}_{r_h}\Big(\frac{1}{2\delta_\epsilon}\big(\nabla_y g(x_t,y_t+\delta_\epsilon v_t)-\nabla_y g(x_t,y_t)\\
&\quad+\nabla_y g(x_t,y_t)-\nabla_y g(x_t,y_t-\delta_\epsilon v_t)\big)\Big)\\
&=\mathcal{M}_{r_h}\Big(\frac{1}{2\delta_\epsilon}\int_{k=0}^1\nabla^2_{yy}g(x_t,y_t+k\delta_\epsilon v_t)\delta_\epsilon v_t dk\\
&\quad+\frac{1}{2\delta_\epsilon}\int_{k=0}^1\nabla^2_{yy}g(x_t,y_t-k\delta_\epsilon v_t)\delta_\epsilon v_t dk\Big)\\
&=\mathcal{M}_{r_h}\Big(\Big(\frac{1}{2}\int_{k=0}^1\nabla^2_{yy}g(x_t,y_t+k\delta_\epsilon v_t)dk\\
&\quad+\frac{1}{2}\int_{k=0}^1\nabla^2_{yy}g(x_t,y_t-k\delta_\epsilon v_t)dk\Big)v_t\Big)\\
&=\mathcal{S}_{[\mu,L_g]}\Big[\frac{1}{2}\int_{k=0}^1\nabla^2_{yy}g(x_t,y_t+k\delta_\epsilon v_t)dk\\
&\quad+\frac{1}{2}\int_{k=0}^1\nabla^2_{yy}g(x_t,y_t-k\delta_\epsilon v_t)dk\Big)\Big]v_t,
\end{aligned}\tag{16}$$

where the last equality holds by $v_t=\mathcal{P}_{r_v}(v_t)$ and the above definition 3.1. Thus, we have

$$\begin{aligned}
&\lim_{\delta_\epsilon\to0}\mathcal{M}_{r_h}\big(\widetilde{H}(x_t,y_t,v_t,\delta_\epsilon)\big)\\
&=\lim_{\delta_\epsilon\to0}\mathcal{S}_{[\mu,L_g]}\Big[\frac{1}{2}\int_{k=0}^1\nabla^2_{yy}g(x_t,y_t+k\delta_\epsilon v_t)dk\\
&\quad+\frac{1}{2}\int_{k=0}^1\nabla^2_{yy}g(x_t,y_t-k\delta_\epsilon v_t)dk\Big)\Big]v_t\\
&=\mathcal{S}_{[\mu,L_g]}[\nabla^2_{yy}g(x_t,y_t)]v_t.
\end{aligned}\tag{17}$$

Then we can obtain $\lim_{\delta_\epsilon\to0}\widetilde{\nabla}_v R(x_t,y_t,v_t)=\nabla_v R(x_t,y_t,v_t)$.

4. Convergence Analysis

In the section, we study convergence properties of our HJFBiO algorithm under some mild assumptions. Given x_t from Algorithm 1, we define a useful gradient mapping

$$\mathcal{G}(x_t,\nabla F(x_t),\gamma)=\frac{1}{\gamma}(x_t-x_{t+1}),\tag{18}$$

where $F(x)=f(x,y^*(x))$ with $y^*(x)\in\arg\min_y g(x,y)$, and x_{t+1} is generated from

$$\begin{aligned}
x_{t+1}&=\mathbb{P}^\gamma_{\phi(\cdot)}(x_t,\nabla F(x_t))\\
&=\arg\min_{x\in\mathbb{R}^d}\Big\{\langle\nabla F(x_t),x\rangle+\frac{1}{2\gamma}\|x-x_t\|^2+\phi(x)\Big\}.
\end{aligned}$$

When $\phi(x)\equiv0$, based on (18), we have $x_{t+1}=x_t-\gamma\nabla F(x_t)$, and then we obtain $\mathcal{G}(x_t,\nabla F(x_t),\gamma)=\nabla F(x_t)$.

4.1. Convergence Properties of Our Algorithm on Unimodal $g(x,y)$

In the subsection, we study the convergence properties of our HJFBiO algorithm when $g(x,\cdot)$ satisfies the global PL condition for all $x\in\mathbb{R}^d$, i.e., it has **a unique minimizer** $y^*(x)=\arg\min_y g(x,y)$. We first give three useful lemmas.

Lemma 4.1. *Suppose the sequence $\{x_t,y_t,v_t\}_{t=1}^T$ be generated from Algorithm 1. Under the above Assumptions, given $0<\tau\le\frac{1}{6L_g}$, we have*

$$\begin{aligned}
\|v_{t+1}-v^*_{t+1}\|^2&\le(1-\frac{\mu\tau}{4})\|v_t-v^*_t\|^2-\frac{3}{4}\|v_{t+1}-v_t\|^2\\
&\quad+\frac{25\tau L^2_{gyy}r_v^4\delta_\epsilon^2}{6\mu}+\frac{20}{3}\Big(\frac{L_f^2}{\mu^2}+\frac{L^2_{gyy}C^2_{fx}}{\mu^4}\Big)\Big(\|x_{t+1}-x_t\|^2\\
&\quad+\|y_{t+1}-y_t\|^2\Big),
\end{aligned}\tag{19}$$

where $v^*_t=v^*(x_t,y_t)=\Big(\mathcal{S}_{[\mu,L_g]}[\nabla^2_{yy}g(x_t,y_t)]\Big)^{-1}\nabla_y f(x_t,y_t)$ for all $t\ge1$.

Lemma 4.2. *Assume the sequence $\{x_t,y_t,v_t\}_{t=1}^T$ be generated from Algorithm 1, given $0<\gamma\le\frac{1}{2L_F}$, we have*

$$\begin{aligned}
\Phi(x_{t+1})&\le\Phi(x_t)-\frac{\gamma}{2}\|\mathcal{G}(x_t,w_t,\gamma)\|^2\\
&\quad+\frac{12\gamma}{\mu}\big(L_f^2+r_v^2L^2_{gxy}\big)\big(g(x_t,y_t)-G(x_t)\big)\\
&\quad+6\gamma C^2_{gxy}\|v_t-v^*_t\|^2+2\gamma L^2_{gxy}\delta_\epsilon^2r_v^4,
\end{aligned}$$

where $\Phi(x)=F(x)+\phi(x)$ and $\mathcal{G}(x_t,w_t,\gamma)=\frac{1}{\gamma}(x_t-x_{t+1})$.

Lemma 4.3. *Suppose the sequence $\{x_t,y_t,v_t\}_{t=1}^T$ be generated from Algorithm 1. Under the above Assumptions 2.1-2.3, given $\gamma\le\min\Big\{\frac{\lambda\mu}{16L_G},\frac{\mu}{16L_g^2}\Big\}$ and $0<\lambda\le\frac{1}{2L_g}$, we have*

$$\begin{aligned}
&g(x_{t+1},y_{t+1})-G(x_{t+1})\\
&\le(1-\frac{\lambda\mu}{2})\big(g(x_t,y_t)-G(x_t)\big)+\frac{1}{8\gamma}\|x_{t+1}-x_t\|^2\\
&\quad-\frac{1}{4\lambda}\|y_{t+1}-y_t\|^2+\lambda\|\nabla_y g(x_t,y_t)-u_t\|^2,
\end{aligned}\tag{20}$$

where $G(x_t) = g(x_t, y^*(x_t))$ *with* $y^*(x_t) \in \arg\min_y g(x_t, y)$ *for all* $t \geq 1$.

Based on the above useful lemmas, we give the convergence properties of our HJFBiO method in the following.

Theorem 4.4. *Assume the sequence* $\{x_t, y_t, v_t\}_{t=1}^T$ *be generated from our Algorithm 1. Under the above Assumptions 2.1-2.5, let* $0 < \gamma \leq \min\Big(\frac{1}{2L_F}, \frac{\lambda\mu}{16L_G}, \frac{\mu}{16L_g^2}, \frac{3}{160\breve{L}^2}, \frac{\mu\tau}{30C_{gxy}^2}, \frac{\mu^2\lambda}{30(L_f^2+r_v^2L_{gxy}^2)}\Big)$, $0 < \lambda \leq \min\big(\frac{1}{2L_g}, \frac{3}{80\breve{L}^2}\big)$ *and* $0 < \tau \leq \frac{1}{6L_g}$, *we have*

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^T \|\mathcal{G}(x_t, \nabla F(x_t), \gamma)\|^2 \\ &\leq \frac{8(\Phi(x_1) + g(x_1, y_1) - G(x_1) + \|v_1 - v_1^*\|^2 - \Phi^*)}{T\gamma} \\ &\quad + 20L_{gxy}^2\delta_\epsilon^2 r_v^4 + \frac{100\tau L_{gyy}^2 r_v^4 \delta_\epsilon^2}{3\gamma\mu}, \end{aligned} \tag{21}$$

where $\breve{L}^2 = \frac{L_f^2}{\mu^2} + \frac{L_{gyy}^2 C_{fx}^2}{\mu^4}$.

Remark 4.5. Without loss of generality, let $\tau = O(1)$, $\lambda = O(1)$ and $\gamma = \min\big(\frac{1}{2L_F}, \frac{\lambda\mu}{16L_G}, \frac{\mu}{16L_g^2}, \frac{3}{160\breve{L}^2}, \frac{\mu\tau}{30C_{gxy}^2}, \frac{\mu^2\lambda}{30(L_f^2+r_v^2L_{gxy}^2)}\big) = O(1)$. Furthermore, let $\delta_\epsilon = O\big(\frac{1}{\sqrt{T}\max(L_{gxy}^2, L_{gyy}^2/\mu)r_v^2}\big)$, we have

$$\frac{1}{T}\sum_{t=1}^T \|\mathcal{G}(x_t, \nabla F(x_t), \gamma)\|^2 \leq O(\frac{1}{T}).$$

Let $O(\frac{1}{T}) = \epsilon$, we obtain $T = (\epsilon^{-1})$. Since requiring seven first-order gradients at each iteration, our HJFBiO algorithm can obtain an optimal gradient (or iteration) complexity of $7 \cdot T = O(\epsilon^{-1})$ in finding an ϵ-stationary solution of Problem (1), which matches the lower bound established by the first-order method for finding an ϵ-stationary point of nonconvex smooth optimization (Carmon et al., 2020).

When $\phi(x) \equiv 0$, based on (18), we have $\mathcal{G}(x_t, \nabla F(x_t), \gamma) = \nabla F(x_t)$. Thus our HJFBiO algorithm still obtains an optimal gradient complexity of $7 \cdot T = O(\epsilon^{-1})$ in finding an ϵ-stationary solution of Problem (1) (i.e., $\min_{1\leq t\leq T} \|\nabla F(x_t)\|^2 \leq \epsilon$).

4.2. Convergence Properties of Our Algorithm on multimodal $g(x, y)$

In this subsection, we study the convergence properties of our HJFBiO method when $g(x, \cdot)$ satisfies the local PL condition for all x, i.e, it has **multi local minimizers** $y^\diamond(x, y) \in \arg\min_y g(x, y)$. As in (Liu et al., 2022), we define the attraction points.

Definition 4.6. Given any (x, y), if sequence $\{y_t\}_{t=0}^\infty$ generated by gradient descent $y_t = y_{t-1} - \lambda\nabla_y g(x, y_{t-1})$ starting from $y_0 = y$ converges to $y^\diamond(x, y)$, we say that $y^\diamond(x, y)$ is the attraction point of (x, y) with step size $\lambda > 0$.

An attraction basin be formed by the same attraction point in set of (x, y). In the following analysis, we assume the PL condition within the individual attraction basins. Let $F^\diamond(x) = f(x, y^\diamond(x, y))$.

Assumption 4.7. (Local PL Condition in Attraction Basins) Assume that for any (x, y), $y^\diamond(x, y)$ exists. $g(x, \cdot)$ satisfies the local PL condition in attraction basins, i.e., for any (x, y), there exists a constant $\mu > 0$ such that

$$\|\nabla_y g(x, y)\|^2 \geq 2\mu\big(g(x, y) - G^\diamond(x)\big), \tag{22}$$

where $G^\diamond(x) = g(x, y^\diamond(x, y))$.

Assumption 4.8. The function $g\big(x, y^\diamond(x, y)\big)$ satisfies

$$\varrho\big(\nabla_{yy}^2 g\big(x, y^\diamond(x, y)\big)\big) \in [\mu, L_g], \tag{23}$$

where $y^\diamond(x, y)$ is the attraction point of (x, y), and $\varrho(\cdot)$ denotes the eigenvalue (or singular-value) function and $L_g \geq \mu > 0$.

Assumption 4.9. The function $\Phi^\diamond(x) = F^\diamond(x) + \phi(x)$ is bounded below in $\mathbb{R}^d$, *i.e.,* $\Phi^\diamond = \inf_{x\in\mathbb{R}^d} \Phi^\diamond(x) > -\infty$.

Theorem 4.10. *Assume the sequence* $\{x_t, y_t, v_t\}_{t=1}^T$ *be generated from our Algorithm 1. Under the above Assumptions 4.7, 4.8, 2.3, 2.4, 4.9, let* $0 < \gamma \leq \min\Big(\frac{1}{2L_F}, \frac{\lambda\mu}{16L_G}, \frac{\mu}{16L_g^2}, \frac{3}{160\breve{L}^2}, \frac{\mu\tau}{30C_{gxy}^2}, \frac{\mu^2\lambda}{30(L_f^2+r_v^2L_{gxy}^2)}\Big)$, $0 < \lambda \leq \min\big(\frac{1}{2L_g}, \frac{3}{80\breve{L}^2}\big)$ *and* $0 < \tau \leq \frac{1}{6L_g}$, *we have*

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^T \|\mathcal{G}(x_t, \nabla F^\diamond(x_t), \gamma)\|^2 \\ &\leq \frac{8(\Phi^\diamond(x_1) + g(x_1, y_1) - G(x_1) + \|v_1 - v_1^*\|^2 - \Phi^\diamond)}{T\gamma} \\ &\quad + 20L_{gxy}^2\delta_\epsilon^2 r_v^4 + \frac{100\tau L_{gyy}^2 r_v^4 \delta_\epsilon^2}{3\gamma\mu}, \end{aligned} \tag{24}$$

where $\Phi^\diamond(x) = F^\diamond(x) + \phi(x)$ *and* $\breve{L}^2 = \frac{L_f^2}{\mu^2} + \frac{L_{gyy}^2 C_{fx}^2}{\mu^4}$.

Remark 4.11. The proof of Theorem 4.10 **can follow** the proof of Theorem 4.4. Let further $\delta_\epsilon = O\big(\frac{1}{\sqrt{T}\max(L_{gxy}^2, L_{gyy}^2/\mu)r_v^2}\big)$, our HJFBiO algorithm can also obtain an optimal gradient (or iteration) complexity of $7 \cdot T = O(\epsilon^{-1})$ in finding an ϵ-stationary solution of Problem (1) **under local PL condition**.

5. Experiments

In the section, we conduct bilevel PL game and hyper-representation learning tasks to demonstrate efficiency of our method. In the experiments, we compare our method with the existing methods given in Table 1. Meanwhile, we add a baseline: BVFSM (Liu et al., 2021b). For fair comparison, since only the AdaPAG (Huang, 2023a) uses adaptive learning rate, we exclude it in the comparisons.

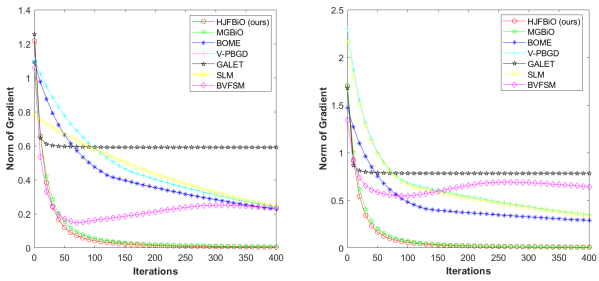

Figure 1: PL Game: norm of gradient vs number of iteration under $d = 100$ (Left) and $d = 200$ (Right).

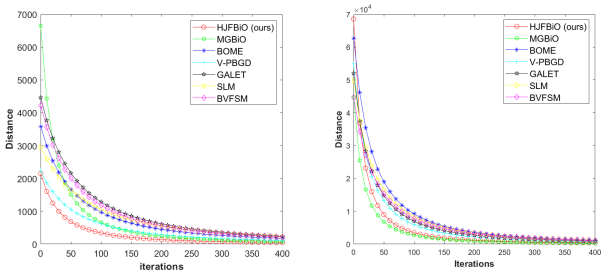

Figure 2: Distances of the algorithms under the case of $d = 100$ (Left) and $d = 200$ (Right).

5.1. Bilevel Polyak-Łojasiewicz Game

In this subsection, as in (Huang, 2023b), we apply the bilevel Polyak-Łojasiewicz game task to verify efficiency of our algorithm, defined as:

$$\begin{aligned} &\min_{x\in\mathbb{R}^d} \frac{1}{2}x^T P x + x^T R^1 y, \\ &\text{s.t. } \min_{y\in\mathbb{R}^d} \frac{1}{2}y^T Q y + x^T R^2 y, \end{aligned} \tag{25}$$

where $P = \frac{1}{n}\sum_{i=1}^n p_i(p_i)^T$, $Q = \frac{1}{n}\sum_{i=1}^n q_i(q_i)^T$, $R^1 = \frac{1}{n}\sum_{i=1}^n 0.01 r_i^1(r_i^1)^T$ and $R^2 = \frac{1}{n}\sum_{i=1}^n 0.01 r_i^2(r_i^2)^T$. Specifically, samples $\{p_i\}_{i=1}^n$, $\{q_i\}_{i=1}^n$, $\{r_i^1\}_{i=1}^n$ and $\{r_i^2\}_{i=1}^n$ are independently drawn from normal distributions $\mathbb{N}(0,\Sigma_P)$, $\mathbb{N}(0,\Sigma_Q)$, $\mathbb{N}(0,\Sigma_{R^1})$ and $\mathbb{N}(0,\Sigma_{R^2})$, respectively. Here we set $\Sigma_P = U^1 D^1 (U^1)^T$, where $U^1 \in \mathbb{R}^{d\times l}$ $(l < d)$ is column orthogonal, and $D^1 \in \mathbb{R}^{l\times l}$ is diagonal and its diagonal elements are distributed uniformly in the interval $[\mu, L]$ with $0 < \mu < L$. Let $\Sigma_Q = U^2 D^2 (U^2)^T$, where $U^2 \in \mathbb{R}^{d\times l}$ is column orthogonal, and $D^2 \in \mathbb{R}^{l\times l}$ is diagonal and its diagonal elements are distributed uniformly in the interval $[\mu, L]$ with $0 < \mu < L$. We also set $\Sigma_{R^1} = 0.001 V^1 (V^1)^T$ and $\Sigma_{R^2} = 0.001 V^2 (V^2)^T$, where each element of $V^1, V^2 \in \mathbb{R}^{d\times d}$ is independently sampled from normal distribution $\mathbb{N}(0,1)$. Since the covariance matrices Σ_P and Σ_Q are rank-deficient, it is ensured that both P and Q are singular.

In the experiment, we set $l = 50$, $n = 50 \cdot d$, $d = 100$ and $d = 200$. For fair comparison, we set a basic learning rate as 0.01 for all algorithms. In our HJFBiO method, we set $\delta_\epsilon = 10^{-5}$. Figure 1 provides the results on *norm of gradient* vs *iteration*, where the iteration denotes iteration at outer loop in all algorithms. Here these results verify that our HJFBiO algorithm outperforms all comparisons. In particular, our HJFBiO method has a lower computation at each iteration than the MGBiO method.

5.2. Hyper-Representation Learning

In this subsection, as in (Huang, 2023b), we consider the hyper-representation learning on matrix sensing task to verify efficiency of our method. Specifically, given n sensing matrices $\{C_i \in \mathbb{R}^{d\times d}\}_{i=1}^n$ and n observations $o_i = \langle C_i, H^*\rangle = \text{trace}(C_i^T H^*)$, where $H^* = U^*(U^*)^T$ is a low-rank symmetric matrix with $U^* \in \mathbb{R}^{d\times r}$, the goal of this task is to find an optimal matrix U^*, which can be defined as the following problem:

$$\min_{U\in\mathbb{R}^{d\times r}} \frac{1}{n}\sum_{i=1}^n \ell_i(U) \equiv \frac{1}{2}\big(\langle C_i, UU^T\rangle - o_i\big)^2. \tag{26}$$

Next, we consider the hyper-representation learning in matrix sensing task, which be rewritten the following bilevel optimization problem:

$$\begin{aligned} &\min_{x\in\mathbb{R}^{d\times r-1}} \frac{1}{|D_v|}\sum_{i\in D_v} \ell_i(x, y^*(x)), \\ &\text{s.t. } y^*(x) \in \arg\min_{y\in\mathbb{R}^d} \frac{1}{|D_t|}\sum_{i\in D_t} \ell_i(x,y), \end{aligned} \tag{27}$$

where $U = [y; x] \in \mathbb{R}^{d\times r}$ is a concatenation of x and y. Here we define variable x to be the first $r-1$ columns of U and variable y to be the last column. Meanwhile, D_t denotes the training dataset, and D_v denotes the validation dataset. The ground truth low-rank matrix H^* is generated by $H^* = U^*(U^*)^T$, where each entry of U^* is drawn from normal distribution $\mathbb{N}(0, 1/d)$ independently. We randomly generate $n = 30 \cdot d$ samples of sensing matrices $\{C_i\}_{i=1}^n$ from standard normal distribution, and then compute the corresponding no-noise labels $o_i = \langle C_i, H^*\rangle$. We split all samples into two dataset: a train dataset D_t with 40% data and a validation dataset D_v with 60% data.

In the experiment, for fair comparison, we set the basic learning rate as 0.01 for all algorithms. In our HJFBiO method, we set $\delta_\epsilon = 10^{-5}$. Let $\ell(U) = \frac{1}{2n}\sum_{i=1}^n \big(\langle C_i, UU^T\rangle - o_i\big)^2$ denote the loss, and $\|UU^T - H^*\|_F^2/\|H^*\|_F^2$ denotes the distance. Figures 2 and 3 show that our HJFBiO method

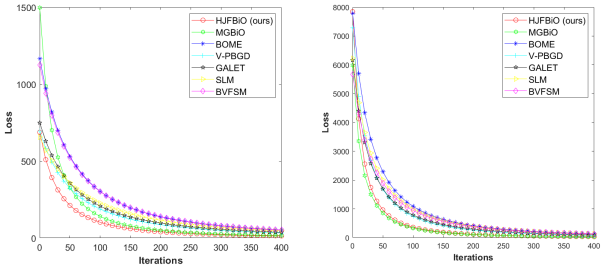

Figure 3: Losses of the algorithms under the case of $d = 100$ (Left) and $d = 200$ (Right).

outperforms all the comparisons on the **distance** vs iteration, where the iteration denotes iteration at outer loop in all algorithms. While our HJFBiO method is comparable with the MGBiO method on the **loss** vs iteration. In particular, our HJFBiO method has a lower computation at each iteration than the MGBiO method.

6. Conclusions

In the paper, we proposed an efficient Hessian/Jacobian-free bilevel method to solve the nonconvex-PL bilevel problems based on the finite-difference estimator and a new projection operator. Moreover, under some mild assumptions, we proved that our HJFBiO method obtains the best known convergence rate $O(\frac{1}{T})$, and gets an optimal gradient (or iteration) complexity of $O(\epsilon^{-1})$ in finding ϵ-stationary solution under global and local PL condition, respectively.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This paper was partially supported by NSFC under Grant No. 62376125. It was also partially supported by the Fundamental Research Funds for the Central Universities NO.NJ2023032.

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

Cao, J., Jiang, R., Abolfazli, N., Hamedani, E. Y., and Mokhtari, A. Projection-free methods for stochastic simple bilevel optimization with convex lower-level problem. *arXiv preprint arXiv:2308.07536*, 2023.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1-2):71–120, 2020.

Chakraborty, S., Bedi, A. S., Koppel, A., Manocha, D., Wang, H., Huang, F., and Wang, M. Aligning agent policy with externalities: Reward design via bilevel rl. *arXiv preprint arXiv:2308.02585*, 2023.

Chen, L., Xu, J., and Zhang, J. On bilevel optimization without lower-level strong convexity. *arXiv preprint arXiv:2301.00712*, 2023a.

Chen, L., Xu, J., and Zhang, J. On finding small hypergradients in bilevel optimization: Hardness results and improved analysis. *37th Annual Conference on Learning Theory*, 2024.

Chen, T., Sun, Y., Xiao, Q., and Yin, W. A single-timescale method for stochastic bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2466–2488. PMLR, 2022.

Chen, Z., Kailkhura, B., and Zhou, Y. An accelerated proximal algorithm for regularized nonconvex and nonsmooth bi-level optimization. *Machine Learning*, 112(5):1433–1463, 2023b.

Colson, B., Marcotte, P., and Savard, G. An overview of bilevel optimization. *Annals of operations research*, 153 (1):235–256, 2007.

Dagréou, M., Ablin, P., Vaiter, S., and Moreau, T. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. In *Advances in Neural Information Processing Systems*, 2022.

Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., and Pontil, M. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pp. 1568–1577. PMLR, 2018.

Frei, S. and Gu, Q. Proxy convexity: A unified framework for the analysis of neural networks trained by gradient descent. *Advances in Neural Information Processing Systems*, 34:7937–7949, 2021.

Ghadimi, S. and Wang, M. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*, 2020.

Huang, F. Adaptive mirror descent bilevel optimization. *arXiv preprint arXiv:2311.04520*, 2023a.

Huang, F. On momentum-based gradient methods for bilevel optimization with nonconvex lower-level. *arXiv preprint arXiv:2303.03944*, 2023b.

Huang, F., Li, J., Gao, S., and Huang, H. Enhanced bilevel optimization via bregman distance. *Advances in Neural Information Processing Systems*, 35:28928–28939, 2022.

Ji, K., Yang, J., and Liang, Y. Bilevel optimization: Convergence analysis and enhanced design. In *International conference on machine learning*, pp. 4882–4892. PMLR, 2021.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016*, pp. 795–811. Springer, 2016.

Kwon, J., Kwon, D., Wright, S., and Nowak, R. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. *arXiv preprint arXiv:2309.01753*, 2023.

Li, J., Gu, B., and Huang, H. A fully single loop algorithm for bilevel optimization without hessian inverse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7426–7434, 2022.

Liu, B., Ye, M., Wright, S., Stone, P., et al. Bome! bilevel optimization made easy: A simple first-order approach. In *Advances in Neural Information Processing Systems*, 2022.

Liu, R., Gao, J., Zhang, J., Meng, D., and Lin, Z. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021a.

Liu, R., Liu, X., Zeng, S., Zhang, J., and Zhang, Y. Value-function-based sequential minimization for bi-level optimization. *arXiv preprint arXiv:2110.04974*, 2021b.

Liu, R., Liu, Y., Yao, W., Zeng, S., and Zhang, J. Averaged method of multipliers for bi-level optimization without lower-level strong convexity. *arXiv preprint arXiv:2302.03407*, 2023.

Lu, S. Slm: A smoothed first-order lagrangian method for structured constrained nonconvex optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Lu, Z. and Mei, S. First-order penalty methods for bilevel optimization. *arXiv preprint arXiv:2301.01716*, 2023.

Nesterov, Y. *Lectures on convex optimization*, volume 137. Springer, 2018.

Polyak, B. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Shen, H. and Chen, T. On penalty-based bilevel gradient descent method. *arXiv preprint arXiv:2302.05185*, 2023.

Song, C., Ramezani-Kebrya, A., Pethick, T., Eftekhari, A., and Cevher, V. Subquadratic overparameterization for shallow neural networks. *Advances in Neural Information Processing Systems*, 34:11247–11259, 2021.

Sow, D., Ji, K., Guan, Z., and Liang, Y. A constrained optimization approach to bilevel optimization with multiple inner minima. *arXiv preprint arXiv:2203.01123*, 2022a.

Sow, D., Ji, K., and Liang, Y. On the convergence theory for hessian-free bilevel algorithms. *Advances in Neural Information Processing Systems*, 35:4136–4149, 2022b.

Xiao, Q., Lu, S., and Chen, T. A generalized alternating method for bilevel optimization under the polyak-{\L} ojasiewicz condition. *arXiv preprint arXiv:2306.02422*, 2023.

Yang, H., Luo, L., Li, C. J., Jordan, M., and Fazel, M. Accelerating inexact hypergradient descent for bilevel optimization. In *OPT 2023: Optimization for Machine Learning*, 2023a.

Yang, Y., Xiao, P., and Ji, K. Achieving $o(\epsilon^{-1.5})$ complexity in hessian/jacobian-free stochastic bilevel optimization. *arXiv preprint arXiv:2312.03807*, 2023b.

A. Detailed Convergence Analysis

In this section, we provide the detailed convergence analysis of our algorithms. We first review some useful lemmas.

Lemma A.1. *((Nesterov, 2018)) Assume that $f(x)$ is a differentiable convex function and $\mathcal{X}$ is a convex set. $x^* \in \mathcal{X}$ is the solution of the constrained problem $\min_{x\in\mathcal{X}} f(x)$, if*

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0, \ \forall x \in \mathcal{X}. \tag{28}$$

Lemma A.2. *((Karimi et al., 2016)) The function $f(x) : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and satisfies PL condition with constant μ, then it also satisfies error bound (EB) condition with μ, i.e., for all $x \in \mathbb{R}^d$*

$$\|\nabla f(x)\| \geq \mu \|x^* - x\|, \tag{29}$$

where $x^* \in \arg\min_x f(x)$. It also satisfies quadratic growth (QG) condition with μ, i.e.,

$$f(x) - \min_x f(x) \geq \frac{\mu}{2}\|x^* - x\|^2. \tag{30}$$

A.1. Convergence Analysis of HJFBiO Algorithm for Bilevel Optimization with Regularization

In this subsection, we detail the convergence analysis of our HJFBiO algorithm for bilevel optimization. We give some useful lemmas.

Lemma A.3. *(Restatement of Lemma 4.1) Suppose the sequence $\{x_t, y_t, v_t\}_{t=1}^T$ be generated from Algorithm 1. Under the above Assumptions 2.1-2.4, given $0 < \tau \leq \frac{1}{6L_g}$, we have*

$$\begin{aligned}\|v_{t+1} - v_{t+1}^*\|^2 \leq (1 - \frac{\mu\tau}{4})\|v_t - v_t^*\|^2 - \frac{3}{4}\|v_{t+1} - v_t\|^2 \\ + \frac{25\tau L_{gyy}^2 r_v^4 \delta_\epsilon^2}{6\mu} + \frac{20}{3}\big(\frac{L_f^2}{\mu^2} + \frac{L_{gyy}^2 C_{fx}^2}{\mu^4}\big)\big(\|x_{t+1} - x_t\|^2 + \|y_{t+1} - y_t\|^2\big),\end{aligned} \tag{31}$$

where $v_t^* = v^*(x_t, y_t) = \Big(\mathcal{S}_{[\mu,L_g]}\big[\nabla_{yy}^2 g(x_t, y_t)\big]\Big)^{-1} \nabla_y f(x_t, y_t)$ for all $t \geq 1$.

Proof. Since the function $R(x, y, v) = \frac{1}{2} v^T \mathcal{S}_{[\mu,L_g]}\big[\nabla_{yy}^2 g(x, y)\big] v - v^T \nabla_y f(x, y)$ is μ-strongly convex on variable v, we have

$$\begin{aligned}R(x_t, y_t, v) &\geq R(x_t, y_t, v_t) + \langle \nabla_v R(x_t, y_t, v_t), v - v_t \rangle + \frac{\mu}{2}\|v - v_t\|^2 \\ &= R(x_t, y_t, v_t) + \langle h_t, v - v_{t+1} \rangle + \langle \nabla_v R(x_t, y_t, v_t) - h_t, v - v_{t+1} \rangle \\ &\quad + \langle \nabla_v R(x_t, y_t, v_t), v_{t+1} - v_t \rangle + \frac{\mu}{2}\|v - v_t\|^2.\end{aligned} \tag{32}$$

Since the function $R(x, y, v)$ is L_g-smooth on variable v, we have

$$R(x_t, y_t, v_{t+1}) \leq R(x_t, y_t, v_t) + \langle \nabla_v R(x_t, y_t, v_t), v_{t+1} - v_t \rangle + \frac{L_g}{2}\|v_{t+1} - v_t\|^2. \tag{33}$$

By combining the about inequalities (32) with (33), we have

$$\begin{aligned}R(x_t, y_t, v) \geq R(x_t, y_t, v_{t+1}) + \langle h_t, v - v_{t+1} \rangle + \langle \nabla_v R(x_t, y_t, v_t) - h_t, v - v_{t+1} \rangle \\ + \frac{\mu}{2}\|v - v_t\|^2 - \frac{L_g}{2}\|v_{t+1} - v_t\|^2.\end{aligned} \tag{34}$$

According to the line 5 of Algorithm 1, we have $v_{t+1} = \mathcal{P}_{r_v}\big(v_t - \tau h_t\big) = \arg\min_{v\in\Lambda_v} \Big\{ \langle h_t, v - v_t \rangle + \frac{1}{2\tau}\|v - v_t\|^2 \Big\}$ with $\Lambda_v = \{v \in \mathbb{R}^p \,|\, \|v\| \leq r_v\}$. According to the above Lemma A.1, given $v \in \Lambda_v$, we have

$$\langle h_t + \frac{1}{\tau}(v_{t+1} - v_t), v - v_{t+1} \rangle \geq 0. \tag{35}$$

By plugging the inequalities (35) into (34), we have

$$R(x_t,y_t,v) \geq R(x_t,y_t,v_{t+1}) + \frac{1}{\tau}\langle v_{t+1}-v_t, v_t - v\rangle + \frac{1}{\tau}\|v_{t+1}-v_t\|^2 \\ + \langle \nabla_v R(x_t,y_t,v_t) - h_t, v - v_{t+1}\rangle + \frac{\mu}{2}\|v-v_t\|^2 - \frac{L_g}{2}\|v_{t+1}-v_t\|^2. \tag{36}$$

Let $v = v_t^* = v^*(x_t,y_t) = \Big(\mathcal{S}_{[\mu,L_g]}\big[\nabla^2_{yy} g(x_t,y_t)\big]\Big)^{-1}\nabla_y f(x_t,y_t)$, then we have

$$R(x_t,y_t,v_t^*) \geq R(x_t,y_t,v_{t+1}) + \frac{1}{\tau}\langle v_{t+1}-v_t, v_t - v_t^*\rangle + (\frac{1}{\tau}-\frac{L_g}{2})\|v_{t+1}-v_t\|^2 \\ + \langle \nabla_v R(x_t,y_t,v_t) - h_t, v_t^* - v_{t+1}\rangle + \frac{\mu}{2}\|v_t^*-v_t\|^2. \tag{37}$$

Due to the strongly-convexity of $R(x,y,\cdot)$ and $v_t^* = \arg\min_{v\in\Lambda_v} R(x_t,y_t,v)$ with $r_v = \frac{C_{fy}}{\mu}$, we have $R(x_t,y_t,v_t^*) \leq R(x_t,y_t,v_{t+1})$. Thus, we obtain

$$0 \geq \frac{1}{\tau}\langle v_{t+1}-v_t, v_t - v_t^*\rangle + (\frac{1}{\tau}-\frac{L_g}{2})\|v_{t+1}-v_t\|^2 \\ + \langle \nabla_v R(x_t,y_t,v_t) - h_t, v_t^* - v_{t+1}\rangle + \frac{\mu}{2}\|v_t^*-v_t\|^2. \tag{38}$$

Consider the term $\langle v_{t+1}-v_t, v_t-v_t^*\rangle$, we have

$$\langle v_{t+1}-v_t, v_t-v_t^*\rangle = \frac{1}{2}\|v_{t+1}-v_t^*\|^2 - \frac{1}{2}\|v_t-v_t^*\|^2 - \frac{1}{2}\|v_{t+1}-v_t\|^2. \tag{39}$$

Consider the upper bound of the term $\langle \nabla_v R(x_t,y_t,v_t) - h_t, v_t^* - v_{t+1}\rangle$, we have

$$\begin{aligned}
&\langle \nabla_v R(x_t,y_t,v_t) - h_t, v_t^* - v_{t+1}\rangle \\
&= \langle \nabla_v R(x_t,y_t,v_t) - h_t, v_t^* - v_t\rangle + \langle \nabla_v R(x_t,y_t,v_t) - h_t, v_t - v_{t+1}\rangle \\
&\geq -\frac{1}{\mu}\|\nabla_v R(x_t,y_t,v_t) - h_t\|^2 - \frac{\mu}{4}\|v_t^*-v_t\|^2 - \frac{1}{\mu}\|\nabla_v R(x_t,y_t,v_t) - h_t\|^2 - \frac{\mu}{4}\|v_t-v_{t+1}\|^2 \\
&= -\frac{2}{\mu}\|\nabla_v R(x_t,y_t,v_t) - h_t\|^2 - \frac{\mu}{4}\|v_t^*-v_t\|^2 - \frac{\mu}{4}\|v_t-v_{t+1}\|^2 \\
&\geq -\frac{2L_{gyy}^2 r_v^4 \delta_\epsilon^2}{\mu} - \frac{\mu}{4}\|v_t^*-v_t\|^2 - \frac{\mu}{4}\|v_t-v_{t+1}\|^2,
\end{aligned} \tag{40}$$

where the last inequality holds by the following inequality:

$$
\begin{aligned}
&\|\nabla_v R(x_t,y_t,v_t)-h_t\|^2\\
&=\|\nabla_v R(x_t,y_t,v_t)-\widetilde{\nabla}_v R(x_t,y_t,v_t)\|^2\\
&=\big\|\mathcal{S}_{[\mu,L_g]}\big[\nabla^2_{yy}g(x_t,y_t)\big]v_t-\nabla_y f(x_t,y_t)-\mathcal{M}_{r_h}\big(\widetilde{H}(x_t,y_t,v_t,\delta_\epsilon)\big)+\nabla_y f(x_t,y_t)\big\|^2\\
&=\big\|\mathcal{S}_{[\mu,L_g]}\big[\nabla^2_{yy}g(x_t,y_t)\big]v_t-\mathcal{M}_{r_h}\big(\widetilde{H}(x_t,y_t,v_t,\delta_\epsilon)\big)\big\|^2\\
&=\big\|\mathcal{S}_{[\mu,L_g]}\big[\nabla^2_{yy}g(x_t,y_t)\big]v_t-\mathcal{M}_{r_h}\Big(\frac{1}{2\delta_\epsilon}\big(\nabla_y g(x_t,y_t+\delta_\epsilon v_t)-\nabla_y g(x_t,y_t-\delta_\epsilon v_t)\big)\Big)\big\|^2\\
&=\big\|\mathcal{S}_{[\mu,L_g]}\big[\nabla^2_{yy}g(x_t,y_t)\big]v_t-\mathcal{M}_{r_h}\Big(\frac{1}{2\delta_\epsilon}\big(\nabla_y g(x_t,y_t+\delta_\epsilon v_t)-\nabla_y g(x_t,y_t)+\nabla_y g(x_t,y_t)-\nabla_y g(x_t,y_t-\delta_\epsilon v_t)\big)\Big)\big\|^2\\
&=\big\|\mathcal{S}_{[\mu,L_g]}\big[\nabla^2_{yy}g(x_t,y_t)\big]v_t-\mathcal{M}_{r_h}\Big(\frac{1}{2\delta_\epsilon}\int_{k=0}^1\nabla^2_{yy}g(x_t,y_t+k\delta_\epsilon v_t)\delta_\epsilon v_t dk+\frac{1}{2\delta_\epsilon}\int_{k=0}^1\nabla^2_{yy}g(x_t,y_t-k\delta_\epsilon v_t)\delta_\epsilon v_t dk\Big)\big\|^2\\
&=\big\|\mathcal{S}_{[\mu,L_g]}\big[\nabla^2_{yy}g(x_t,y_t)\big]v_t-\mathcal{M}_{r_h}\Big(\big(\frac{1}{2}\int_{k=0}^1\nabla^2_{yy}g(x_t,y_t+k\delta_\epsilon v_t)dk+\frac{1}{2}\int_{k=0}^1\nabla^2_{yy}g(x_t,y_t-k\delta_\epsilon v_t)dk\big)v_t\Big)\big\|^2\\
&\overset{(i)}{=}\big\|\mathcal{S}_{[\mu,L_g]}\big[\nabla^2_{yy}g(x_t,y_t)\big]v_t-\mathcal{S}_{[\mu,L_g]}\Big[\frac{1}{2}\int_{k=0}^1\nabla^2_{yy}g(x_t,y_t+k\delta_\epsilon v_t)dk+\frac{1}{2}\int_{k=0}^1\nabla^2_{yy}g(x_t,y_t-k\delta_\epsilon v_t)dk\big)\Big]v_t\big\|^2\\
&\overset{(ii)}{\le} r_v^2\big\|\nabla^2_{yy}g(x_t,y_t)-\frac{1}{2}\int_{k=0}^1\nabla^2_{yy}g(x_t,y_t+k\delta_\epsilon v_t)dk-\frac{1}{2}\int_{k=0}^1\nabla^2_{yy}g(x_t,y_t-k\delta_\epsilon v_t)dk\big)\big\|^2\\
&\le\frac{r_v^2}{2}\Big(\int_{k=0}^1\big\|\nabla^2_{yy}g(x_t,y_t)-\nabla^2_{yy}g(x_t,y_t+k\delta_\epsilon v_t)\big\|dk\Big)^2+\frac{r_v^2}{2}\Big(\int_{k=0}^1\big\|\nabla^2_{yy}g(x_t,y_t)-\nabla^2_{yy}g(x_t,y_t-k\delta_\epsilon v_t)\big\|dk\Big)^2\\
&\le L^2_{gyy}r_v^4\delta_\epsilon^2,
\end{aligned}
\tag{41}
$$

where the above equality (i) holds by $v_t=\mathcal{P}_{r_v}(v_t)$ and the above definition 3.1, and the above inequality (ii) holds by $\|v_t\|\le r_v$.

By plugging the inequalities (39) and (40) into (38), we obtain

$$
\begin{aligned}
\frac{1}{2\tau}\|v_{t+1}-v_t^*\|^2&\le(\frac{1}{2\tau}-\frac{\mu}{4})\|v_t-v_t^*\|^2+(\frac{\mu}{4}+\frac{L_g}{2}-\frac{1}{2\tau})\|v_{t+1}-v_t\|^2+\frac{2L^2_{gyy}r_v^4\delta_\epsilon^2}{\mu}\\
&\le(\frac{1}{2\tau}-\frac{\mu}{4})\|v_t-v_t^*\|^2+(\frac{3L_g}{4}-\frac{1}{2\tau})\|v_{t+1}-v_t\|^2+\frac{2L^2_{gyy}r_v^4\delta_\epsilon^2}{\mu}\\
&=(\frac{1}{2\tau}-\frac{\mu}{4})\|v_t-v_t^*\|^2-(\frac{3}{8\tau}+\frac{1}{8\tau}-\frac{3L_g}{4})\|v_{t+1}-v_t\|^2+\frac{2L^2_{gyy}r_v^4\delta_\epsilon^2}{\mu}\\
&\le(\frac{1}{2\tau}-\frac{\mu}{4})\|v_t-v_t^*\|^2-\frac{3}{8\tau}\|v_{t+1}-v_t\|^2+\frac{2L^2_{gyy}r_v^4\delta_\epsilon^2}{\mu},
\end{aligned}
\tag{42}
$$

where the second inequality holds by $L_g\ge\mu$, and the last inequality is due to $0<\tau\le\frac{1}{6L_g}$. It implies that

$$
\|v_{t+1}-v_t^*\|^2\le(1-\frac{\mu\tau}{2})\|v_t-v_t^*\|^2-\frac{3}{4}\|v_{t+1}-v_t\|^2+\frac{4\tau L^2_{gyy}r_v^4\delta_\epsilon^2}{\mu}.
\tag{43}
$$

Since $v_t^*=\Big(\mathcal{S}_{[\mu,L_g]}\big[\nabla^2_{yy}g(x_t,y_t)\big]\Big)^{-1}\nabla_y f(x_t,y_t)=\arg\min_{v\in\Lambda_v}R(x_t,y_t,v)$ with $r_v=\frac{C_{fy}}{\mu}$ and

$v_{t+1}^* = \Big(\mathcal{S}_{[\mu,L_g]}\big[\nabla_{yy}^2 g(x_{t+1},y_{t+1})\big]\Big)^{-1}\nabla_y f(x_{t+1},y_{t+1}) = \arg\min_{v\in\Lambda_v} R(x_{t+1},y_{t+1},v)$, we have

$$\begin{aligned}
\|v_{t+1}^* - v_t^*\|^2 &= \Big\|\Big(\mathcal{S}_{[\mu,L_g]}\big[\nabla_{yy}^2 g(x_{t+1},y_{t+1})\big]\Big)^{-1}\nabla_y f(x_{t+1},y_{t+1}) - \Big(\mathcal{S}_{[\mu,L_g]}\big[\nabla_{yy}^2 g(x_t,y_t)\big]\Big)^{-1}\nabla_y f(x_t,y_t)\Big\|^2 \\
&= \Big\|\Big(\mathcal{S}_{[\mu,L_g]}\big[\nabla_{yy}^2 g(x_{t+1},y_{t+1})\big]\Big)^{-1}\nabla_y f(x_{t+1},y_{t+1}) - \Big(\mathcal{S}_{[\mu,L_g]}\big[\nabla_{yy}^2 g(x_{t+1},y_{t+1})\big]\Big)^{-1}\nabla_y f(x_t,y_t) \\
&\quad + \Big(\mathcal{S}_{[\mu,L_g]}\big[\nabla_{yy}^2 g(x_{t+1},y_{t+1})\big]\Big)^{-1}\nabla_y f(x_t,y_t) - \Big(\mathcal{S}_{[\mu,L_g]}\big[\nabla_{yy}^2 g(x_t,y_t)\big]\Big)^{-1}\nabla_y f(x_t,y_t)\Big\|^2 \\
&\le \big(\frac{4L_f^2}{\mu^2} + \frac{4L_{gyy}^2 C_{fx}^2}{\mu^4}\big)\big(\|x_{t+1}-x_t\|^2 + \|y_{t+1}-y_t\|^2\big). \qquad (44)
\end{aligned}$$

Next, we decompose the term $\|v_{t+1} - v_{t+1}^*\|^2$ as follows:

$$\begin{aligned}
\|v_{t+1} - v_{t+1}^*\|^2 &= \|v_{t+1} - v_t^* + v_t^* - v_{t+1}^*\|^2 \qquad (45)\\
&= \|v_{t+1} - v_t^*\|^2 + 2\langle v_{t+1} - v_t^*, v_t^* - v_{t+1}^*\rangle + \|v_t^* - v_{t+1}^*\|^2 \\
&\le (1+\frac{\mu\tau}{4})\|v_{t+1} - v_t^*\|^2 + (1+\frac{4}{\mu\tau})\|v_t^* - v_{t+1}^*\|^2 \\
&\le (1+\frac{\mu\tau}{4})\|v_{t+1} - v_t^*\|^2 + (1+\frac{4}{\mu\tau})\big(\frac{4L_f^2}{\mu^2} + \frac{4L_{gyy}^2 C_{fx}^2}{\mu^4}\big)\big(\|x_{t+1}-x_t\|^2 + \|y_{t+1}-y_t\|^2\big),
\end{aligned}$$

where the first inequality holds by Cauchy-Schwarz inequality and Young's inequality, and the second inequality is due to the above inequality (44).

By combining the above inequalities (43) and (45), we have

$$\begin{aligned}
\|v_{t+1} - v_{t+1}^*\|^2 &\le (1+\frac{\mu\tau}{4})(1-\frac{\mu\tau}{2})\|v_t - v_t^*\|^2 - (1+\frac{\mu\tau}{4})\frac{3}{4}\|v_{t+1} - v_t\|^2 \\
&\quad + (1+\frac{\mu\tau}{4})\frac{4\tau L_{gyy}^2 r_v^4 \delta_\epsilon^2}{\mu} + (1+\frac{4}{\mu\tau})\big(\frac{4L_f^2}{\mu^2} + \frac{4L_{gyy}^2 C_{fx}^2}{\mu^4}\big)\big(\|x_{t+1}-x_t\|^2 + \|y_{t+1}-y_t\|^2\big).
\end{aligned}$$

Since $0 < \tau \le \frac{1}{6L_g}$ and $L_g \ge \mu$, we have $\tau \le \frac{1}{6L_g} \le \frac{1}{6\mu}$. Then we have

$$\begin{aligned}
(1+\frac{\mu\tau}{4})(1-\frac{\mu\tau}{2}) &= 1 - \frac{\mu\tau}{2} + \frac{\mu\tau}{4} - \frac{\mu^2\tau^2}{8} \le 1 - \frac{\mu\tau}{4}, \\
-(1+\frac{\mu\tau}{4})\frac{3}{4} &\le -\frac{3}{4}, \\
(1+\frac{\mu\tau}{4})\frac{4\tau}{\mu} &\le (1+\frac{1}{24})\frac{4\tau}{\mu} = \frac{25\tau}{6\mu}, \\
1 + \frac{4}{\mu\tau} &\le \frac{5}{3}.
\end{aligned}$$

Thus we have

$$\begin{aligned}
\|v_{t+1} - v_{t+1}^*\|^2 &\le (1-\frac{\mu\tau}{4})\|v_t - v_t^*\|^2 - \frac{3}{4}\|v_{t+1} - v_t\|^2 \\
&\quad + \frac{25\tau L_{gyy}^2 r_v^4 \delta_\epsilon^2}{6\mu} + \frac{20}{3}\big(\frac{L_f^2}{\mu^2} + \frac{L_{gyy}^2 C_{fx}^2}{\mu^4}\big)\big(\|x_{t+1}-x_t\|^2 + \|y_{t+1}-y_t\|^2\big). \qquad (46)
\end{aligned}$$

□

Lemma A.4. *(Restatement of Lemma 4.2) Assume the sequence $\{x_t, y_t, v_t\}_{t=1}^T$ be generated from Algorithm 1, given $0 < \gamma \le \frac{1}{2L_F}$, we have*

$$\begin{aligned}
\Phi(x_{t+1}) &\le \Phi(x_t) - \frac{\gamma}{2}\|\mathcal{G}(x_t,w_t,\gamma)\|^2 + \frac{12\gamma}{\mu}\big(L_f^2 + r_v^2 L_{gxy}^2\big)\big(g(x_t,y_t) - G(x_t)\big) \\
&\quad + 6\gamma C_{gxy}^2\|v_t - v_t^*\|^2 + 2\gamma L_{gxy}^2\delta_\epsilon^2 r_v^4,
\end{aligned}$$

where $\Phi(x) = F(x) + \phi(x)$ and $\mathcal{G}(x_t,w_t,\gamma) = \frac{1}{\gamma}(x_t - x_{t+1})$.

Proof. By the line 4 of Algorithm 1, we have

$$x_{t+1} = \mathbb{P}^{\gamma}_{\phi(\cdot)}\big(x_t, w_t\big) = \arg\min_{x\in\mathbb{R}^d}\Big\{\langle w_t, x\rangle + \frac{1}{2\gamma}\|x - x_t\|^2 + \phi(x)\Big\}. \tag{47}$$

Then we define a gradient mapping $\mathcal{G}(x_t, w_t, \gamma) = \frac{1}{\gamma}(x_t - x_{t+1})$. By the optimality condition of the subproblem (47), we have for any $z \in \mathbb{R}^d$

$$\big\langle w_t + \frac{1}{\gamma}(x_{t+1} - x_t) + \vartheta_{t+1}, z - x_{t+1}\big\rangle \geq 0, \tag{48}$$

where $\vartheta_{t+1} \in \partial\phi(x_{t+1})$.

Let $z = x_t$, and by the convexity of $\phi(x)$, we can obtain

$$\begin{aligned}\langle w_t, x_t - x_{t+1}\rangle &\geq \frac{1}{\gamma}\|x_{t+1} - x_t\|^2 + \langle \vartheta_{t+1}, x_{t+1} - x_t\rangle \\ &\geq \frac{1}{\gamma}\|x_{t+1} - x_t\|^2 + \phi(x_{t+1}) - \phi(x_t).\end{aligned} \tag{49}$$

According to the Lemma 2.7, function $F(x)$ has L_F-Lipschitz continuous gradient. Let $\mathcal{G}(x_t, w_t, \gamma) = \frac{1}{\gamma}(x_t - x_{t+1})$, we have

$$\begin{aligned}F(x_{t+1}) &\leq F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t\rangle + \frac{L_F}{2}\|x_{t+1} - x_t\|^2 \\ &= F(x_t) + \langle w_t, x_{t+1} - x_t\rangle + \gamma\langle \nabla F(x_t) - w_t, \mathcal{G}(x_t, w_t, \gamma)\rangle + \frac{\gamma^2 L_F}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 \\ &\leq F(x_t) - \gamma\|\mathcal{G}(x_t, w_t, \gamma)\|^2 - \phi(x_{t+1}) + \phi(x_t) + \gamma\langle \nabla F(x_t) - w_t, \mathcal{G}(x_t, w_t, \gamma)\rangle + \frac{\gamma^2 L_F}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 \\ &\overset{(ii)}{\leq} F(x_t) - \frac{\gamma}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 - \phi(x_{t+1}) + \phi(x_t) + \gamma\|w_t - \nabla F(x_t)\|^2,\end{aligned} \tag{50}$$

where the second last inequality holds by the above inequality (49), and the last inequality holds by $0 < \gamma \leq \frac{1}{2L_F}$ and the following inequality

$$\begin{aligned}\langle \nabla F(x_t) - w_t, \mathcal{G}(x_t, w_t, \gamma)\rangle &\leq \|w_t - \nabla F(x_t)\|\|\mathcal{G}(x_t, w_t, \gamma)\| \\ &\leq \frac{\rho}{2}\|w_t - \nabla F(x_t)\|^2 + \frac{1}{2\rho}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 \\ &= \|w_t - \nabla F(x_t)\|^2 + \frac{1}{4}\|\mathcal{G}(x_t, w_t, \gamma)\|^2,\end{aligned} \tag{51}$$

where the above inequality holds by Young inequality with $\rho = 2$.

Since $w_t = \widetilde{\nabla} f(x_t, y_t, v_t)$ in Algorithm 1, we have

$$\begin{aligned}\|w_t - \nabla F(x_t)\|^2 &= \|\widetilde{\nabla} f(x_t, y_t, v_t) - \nabla F(x_t)\|^2 \\ &\leq 2\|\widehat{\nabla} f(x_t, y_t, v_t) - \nabla F(x_t)\|^2 + 2\|\widetilde{\nabla} f(x_t, y_t, v_t) - \widehat{\nabla} f(x_t, y_t, v_t)\|^2 \\ &\leq \frac{12}{\mu}\big(L_f^2 + r_v^2 L_{gxy}^2\big)\big(g(x_t, y_t) - G(x_t)\big) + 6C_{gxy}^2\|v_t - v_t^*\|^2 + 2L_{gxy}^2\delta_\epsilon^2 r_v^4,\end{aligned} \tag{52}$$

where the last inequality holds by the following inequalities (53) and (54).

Considering the term $\|\widehat{\nabla} f(x_t, y_t, v_t) - \nabla F(x_t)\|^2$, we have

$$\begin{aligned}
&\|\widehat{\nabla} f(x_t, y_t, v_t) - \nabla F(x_t)\|^2 \\
&= \|\nabla_x f(x_t, y_t) - \nabla^2_{xy} g(x_t, y_t) v_t - \nabla F(x_t)\|^2 \\
&= \|\nabla_x f(x_t, y_t) - \nabla^2_{xy} g(x_t, y_t) v_t - \nabla_x f(x_t, y^*(x_t)) + \nabla^2_{xy} g(x_t, y^*(x_t)) v_t^*\|^2 \\
&= \|\nabla_x f(x_t, y_t) - \nabla_x f(x_t, y^*(x_t)) - \nabla^2_{xy} g(x_t, y_t) v_t + \nabla^2_{xy} g(x_t, y^*(x_t)) v_t \\
&\quad - \nabla^2_{xy} g(x_t, y^*(x_t)) v_t + \nabla^2_{xy} g(x_t, y^*(x_t)) v_t^*\|^2 \\
&\le \big(3L_f^2 + 3r_v^2 L_{gxy}^2\big)\|y_t - y^*(x_t)\|^2 + 3C_{gxy}^2 \|v_t - v_t^*\|^2 \\
&\le \frac{6}{\mu}\big(L_f^2 + r_v^2 L_{gxy}^2\big)\big(g(x_t, y_t) - G(x_t)\big) + 3C_{gxy}^2 \|v_t - v_t^*\|^2,
\end{aligned} \tag{53}$$

where the last inequality holds by the above Lemma A.2.

Considering the term $\|\widetilde{\nabla} f(x_t, y_t, v_t) - \widehat{\nabla} f(x_t, y_t, v_t)\|^2$, we have

$$\begin{aligned}
&\|\widetilde{\nabla} f(x_t, y_t, v_t) - \widehat{\nabla} f(x_t, y_t, v_t)\|^2 \\
&= \|\nabla_x f(x_t, y_t) - \widetilde{J}(x_t, y_t, v_t, \delta_\epsilon) - \nabla_x f(x_t, y_t) + \nabla^2_{xy} g(x_t, y_t) v_t\|^2 \\
&= \left\| \frac{\nabla_x g(x_t, y_t + \delta_\epsilon v_t) - \nabla_x g(x_t, y_t - \delta_\epsilon v_t)}{2\delta_\epsilon} - \nabla^2_{xy} g(x_t, y_t) v_t \right\|^2 \\
&= \frac{1}{4\delta_\epsilon^2}\big\|\nabla_x g(x_t, y_t + \delta_\epsilon v_t) - \nabla_x g(x_t, y_t) - \delta_\epsilon \nabla^2_{xy} g(x_t, y_t) v_t \\
&\qquad + \nabla_x g(x_t, y_t) - \nabla_x g(x_t, y_t - \delta_\epsilon v_t) - \delta_\epsilon \nabla^2_{xy} g(x_t, y_t) v_t\big\|^2 \\
&\le \frac{1}{2\delta_\epsilon^2}\big\|\nabla_x g(x_t, y_t + \delta_\epsilon v_t) - \nabla_x g(x_t, y_t) - \delta_\epsilon \nabla^2_{xy} g(x_t, y_t) v_t\big\|^2 \\
&\quad + \frac{1}{2\delta_\epsilon^2}\big\|\nabla_x g(x_t, y_t) - \nabla_x g(x_t, y_t - \delta_\epsilon v_t) - \delta_\epsilon \nabla^2_{xy} g(x_t, y_t) v_t\big\|^2 \\
&= \frac{1}{2\delta_\epsilon^2}\Big\| \int_{k=0}^1 \big(\nabla^2_{xy} g(x_t, y_t + k\delta_\epsilon v_t) - \nabla^2_{xy} g(x_t, y_t)\big)\delta_\epsilon v_t dk\Big\|^2 \\
&\quad + \frac{1}{2\delta_\epsilon^2}\Big\| \int_{k=0}^1 \big(\nabla^2_{xy} g(x_t, y_t - k\delta_\epsilon v_t) - \nabla^2_{xy} g(x_t, y_t)\big)\delta_\epsilon v_t dk\Big\|^2 \\
&\le \frac{1}{2\delta_\epsilon^2}\Big(\int_{k=0}^1 \big\|\nabla^2_{xy} g(x_t, y_t + k\delta_\epsilon v_t) - \nabla^2_{xy} g(x_t, y_t)\big\| \|v_t\| \delta_\epsilon dk\Big)^2 \\
&\quad + \frac{1}{2\delta_\epsilon^2}\Big(\int_{k=0}^1 \big\|\nabla^2_{xy} g(x_t, y_t - k\delta_\epsilon v_t) - \nabla^2_{xy} g(x_t, y_t)\big\| \|v_t\| \delta_\epsilon dk\Big)^2 \\
&\le L_{gxy}^2 \delta_\epsilon^2 \|v_t\|^4 \le L_{gxy}^2 \delta_\epsilon^2 r_v^4,
\end{aligned} \tag{54}$$

where the last inequality holds by $\|v_t\| \le r_v$.

By combining the above inequalities (50) with (52), we can obtain

$$\begin{aligned}
\Phi(x_{t+1}) \le \Phi(x_t) &- \frac{\gamma}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + \frac{12\gamma}{\mu}\big(L_f^2 + r_v^2 L_{gxy}^2\big)\big(g(x_t, y_t) - G(x_t)\big) \\
&+ 6\gamma C_{gxy}^2 \|v_t - v_t^*\|^2 + 2\gamma L_{gxy}^2 \delta_\epsilon^2 r_v^4.
\end{aligned}$$

□

Lemma A.5. *(Restatement of Lemma 4.3) Suppose the sequence $\{x_t, y_t, v_t\}_{t=1}^T$ be generated from Algorithm 1. Under the*

above Assumptions 2.1-2.3, given $\gamma \leq \min\left\{\frac{\lambda\mu}{16L_G}, \frac{\mu}{16L_g^2}\right\}$ *and* $0 < \lambda \leq \frac{1}{2L_g}$*, we have*

$$g(x_{t+1}, y_{t+1}) - G(x_{t+1}) \leq (1 - \frac{\lambda\mu}{2})\big(g(x_t, y_t) - G(x_t)\big) + \frac{1}{8\gamma}\|x_{t+1} - x_t\|^2 - \frac{1}{4\lambda}\|y_{t+1} - y_t\|^2 + \lambda\|\nabla_y g(x_t, y_t) - u_t\|^2, \tag{55}$$

where $G(x_t) = g(x_t, y^*(x_t))$ *with* $y^*(x_t) \in \arg\min_y g(x_t, y)$ *for all* $t \geq 1$.

Proof. Using the Assumption 2.3, i.e., L_g-smoothness of $g(x, \cdot)$, such that

$$g(x_{t+1}, y_{t+1}) \leq g(x_{t+1}, y_t) + \langle \nabla_y g(x_{t+1}, y_t), y_{t+1} - y_t\rangle + \frac{L_g}{2}\|y_{t+1} - y_t\|^2. \tag{56}$$

Since $y_{t+1} = \arg\min_{y\in\mathbb{R}^p}\left\{\langle u_t, y\rangle + \frac{1}{2\lambda}(y - y_t)^T(y - y_t)\right\} = y_t - \lambda u_t$, we can obtain

$$\begin{aligned}
&\langle \nabla_y g(x_{t+1}, y_t), y_{t+1} - y_t\rangle \\
&= -\lambda\langle \nabla_y g(x_{t+1}, y_t), u_t\rangle \\
&= -\frac{\lambda}{2}\Big(\|\nabla_y g(x_{t+1}, y_t)\|^2 + \|u_t\|^2 - \|\nabla_y g(x_{t+1}, y_t) - \nabla_y g(x_t, y_t) + \nabla_y g(x_t, y_t) - u_t\|^2\Big) \\
&\leq -\frac{\lambda}{2}\|\nabla_y g(x_{t+1}, y_t)\|^2 - \frac{1}{2\lambda}\|y_{t+1} - y_t\|^2 + \lambda L_g^2\|x_{t+1} - x_t\|^2 + \lambda\|\nabla_y g(x_t, y_t) - u_t\|^2 \\
&\leq -\lambda\mu\big(g(x_{t+1}, y_t) - G(x_{t+1})\big) - \frac{1}{2\lambda}\|y_{t+1} - y_t\|^2 + \lambda L_g^2\|x_{t+1} - x_t\|^2 + \lambda\|\nabla_y g(x_t, y_t) - u_t\|^2,
\end{aligned} \tag{57}$$

where the last inequality is due to the quadratic growth condition of μ-PL functions, i.e.,

$$\|\nabla_y g(x_{t+1}, y_t)\|^2 \geq 2\mu\big(g(x_{t+1}, y_t) - \min_{y'} g(x_{t+1}, y')\big) = 2\mu\big(g(x_{t+1}, y_t) - G(x_{t+1})\big). \tag{58}$$

Substituting (57) into (56), we have

$$\begin{aligned}
g(x_{t+1}, y_{t+1}) \leq\ & g(x_{t+1}, y_t) - \lambda\mu\big(g(x_{t+1}, y_t) - G(x_{t+1})\big) - \frac{1}{2\lambda}\|y_{t+1} - y_t\|^2 + \lambda L_g^2\|x_{t+1} - x_t\|^2 \\
&+ \lambda\|\nabla_y g(x_t, y_t) - u_t\|^2 + \frac{L_g}{2}\|y_{t+1} - y_t\|^2,
\end{aligned} \tag{59}$$

then rearranging the terms, we can obtain

$$\begin{aligned}
g(x_{t+1}, y_{t+1}) - G(x_{t+1}) \leq\ & (1 - \lambda\mu)\big(g(x_{t+1}, y_t) - G(x_{t+1})\big) - \frac{1}{2\lambda}\|y_{t+1} - y_t\|^2 + \lambda L_g^2\|x_{t+1} - x_t\|^2 \\
&+ \lambda\|\nabla_y g(x_t, y_t) - u_t\|^2 + \frac{L_g}{2}\|y_{t+1} - y_t\|^2.
\end{aligned} \tag{60}$$

Next, using L_g-smoothness of function $f(\cdot, y_t)$, such that

$$g(x_{t+1}, y_t) \leq g(x_t, y_t) + \langle \nabla_x g(x_t, y_t), x_{t+1} - x_t\rangle + \frac{L_g}{2}\|x_{t+1} - x_t\|^2, \tag{61}$$

then we have

$$\begin{aligned}
&g(x_{t+1}, y_t) - g(x_t, y_t) \\
&\leq \langle \nabla_x g(x_t, y_t), x_{t+1} - x_t\rangle + \frac{L_g}{2}\|x_{t+1} - x_t\|^2 \\
&= \langle \nabla_x g(x_t, y_t) - \nabla G(x_t), x_{t+1} - x_t\rangle + \langle \nabla G(x_t), x_{t+1} - x_t\rangle + \frac{L_g}{2}\|x_{t+1} - x_t\|^2 \\
&\leq \frac{1}{8\gamma}\|x_{t+1} - x_t\|^2 + 2\gamma\|\nabla_x g(x_t, y_t) - \nabla G(x_t)\|^2 + \langle \nabla G(x_t), x_{t+1} - x_t\rangle + \frac{L_g}{2}\|x_{t+1} - x_t\|^2 \\
&\leq \frac{1}{8\gamma}\|x_{t+1} - x_t\|^2 + 2L_g^2\gamma\|y_t - y^*(x_t)\|^2 + G(x_{t+1}) - G(x_t) + \frac{L_G}{2}\|x_{t+1} - x_t\|^2 + \frac{L_g}{2}\|x_{t+1} - x_t\|^2 \\
&\leq \frac{4L_g^2\gamma}{\mu}\big(g(x_t, y_t) - G(x_t)\big) + G(x_{t+1}) - G(x_t) + (\frac{1}{8\gamma} + L_G)\|x_{t+1} - x_t\|^2,
\end{aligned} \tag{62}$$

where the second last inequality is due to L_G-smoothness of function $G(x)$, and the last inequality holds by Lemma A.2 and $L_g \le L_G$. Then we have

$$\begin{aligned} g(x_{t+1},y_t) - G(x_{t+1}) &= g(x_{t+1},y_t) - g(x_t,y_t) + g(x_t,y_t) - G(x_t) + G(x_t) - G(x_{t+1}) \\ &\le (1+\frac{4L_g^2\gamma}{\mu})\big(g(x_t,y_t)-G(x_t)\big) + (\frac{1}{8\gamma}+L_G)\|x_{t+1}-x_t\|^2. \end{aligned} \tag{63}$$

Substituting (63) in (60), we get

$$\begin{aligned} &g(x_{t+1},y_{t+1}) - G(x_{t+1}) \\ &\le (1-\lambda\mu)(1+\frac{4L_g^2\gamma}{\mu})\big(g(x_t,y_t)-G(x_t)\big) + (1-\lambda\mu)(\frac{1}{8\gamma}+L_G)\|x_{t+1}-x_t\|^2 \\ &\quad - \frac{1}{2\lambda}\|y_{t+1}-y_t\|^2 + \lambda L_g^2\|x_{t+1}-x_t\|^2 + \lambda\|\nabla_y g(x_t,y_t)-u_t\|^2 + \frac{L_g}{2}\|y_{t+1}-y_t\|^2 \\ &= (1-\lambda\mu)(1+\frac{4L_g^2\gamma}{\mu})\big(g(x_t,y_t)-G(x_t)\big) + \big(\frac{1}{8\gamma}+L_G-\frac{\lambda\mu}{8\gamma}-L_G\lambda\mu+L_g^2\lambda\big)\|x_{t+1}-x_t\|^2 \\ &\quad - \frac{1}{2}\big(\frac{1}{\lambda}-L_g\big)\|y_{t+1}-y_t\|^2 + \lambda\|\nabla_y g(x_t,y_t)-u_t\|^2 \\ &\le (1-\frac{\lambda\mu}{2})\big(g(x_t,y_t)-G(x_t)\big) + \frac{1}{8\gamma}\|x_{t+1}-x_t\|^2 - \frac{1}{4\lambda}\|y_{t+1}-y_t\|^2 + \lambda\|\nabla_y g(x_t,y_t)-u_t\|^2, \end{aligned} \tag{64}$$

where the last inequality holds by $\gamma \le \min\Big\{\frac{\lambda\mu}{16L_G}, \frac{\mu}{16L_g^2}\Big\}$, $L_G \ge L_g(1+\kappa)^2$ and $\lambda \le \frac{1}{2L_g}$ for all $t \ge 1$, i.e.,

$$\begin{aligned} &\gamma \le \frac{\lambda\mu}{16L_G} \Rightarrow \lambda \ge \frac{16L_G\gamma}{\mu} \ge \frac{16L_g}{\mu}(1+\kappa)^2\gamma \ge 8\kappa^2\gamma \Rightarrow \frac{\lambda\mu}{2} \ge \frac{4L_g^2\gamma}{\mu} \\ &\gamma \le \min\Big\{\frac{\lambda\mu}{16L_G}, \frac{\mu}{16L_g^2}\Big\} \Rightarrow \frac{\lambda\mu}{8\gamma} \ge L_G + L_g^2\lambda \\ &\lambda \le \frac{1}{2L_g} \Rightarrow \frac{1}{2\lambda} \ge L_g,\ \forall t \ge 1. \end{aligned} \tag{65}$$

□

Theorem A.6. *(Restatement of Theorem 4.4) Assume the sequence $\{x_t, y_t, v_t\}_{t=1}^T$ be generated from our Algorithm 1. Under the above Assumptions 2.1-2.5, let $0 < \gamma \le \min\Big(\frac{1}{2L_F}, \frac{\lambda\mu}{16L_G}, \frac{\mu}{16L_g^2}, \frac{3}{160\breve{L}^2}, \frac{\mu\tau}{30C_{gxy}^2}, \frac{\mu^2\lambda}{30(L_f^2+r_v^2L_{gxy}^2)}\Big)$, $0 < \lambda \le \min\big(\frac{1}{2L_g}, \frac{3}{80\breve{L}^2}\big)$ and $0 < \tau \le \frac{1}{6L_g}$, we have*

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^T \|\mathcal{G}(x_t,\nabla F(x_t),\gamma)\|^2 \\ &\le \frac{8(\Phi(x_1)+g(x_1,y_1)-G(x_1)+\|v_1-v_1^*\|^2-\Phi^*)}{T\gamma} + 20L_{gxy}^2\delta_\epsilon^2 r_v^4 + \frac{100\tau L_{gyy}^2 r_v^4\delta_\epsilon^2}{3\gamma\mu}, \end{aligned} \tag{66}$$

where $\Phi(x) = F(x)+\phi(x)$ and $\breve{L}^2 = \frac{L_f^2}{\mu^2} + \frac{L_{gyy}^2C_{fx}^2}{\mu^4}$.

Proof. According to Lemma A.4, we have

$$\begin{aligned} \Phi(x_{t+1}) - \Phi(x_t) \le &-\frac{\gamma}{2}\|\mathcal{G}(x_t,w_t,\gamma)\|^2 + \frac{12\gamma}{\mu}\big(L_f^2+r_v^2L_{gxy}^2\big)\big(g(x_t,y_t)-G(x_t)\big) \\ &+ 6\gamma C_{gxy}^2\|v_t-v_t^*\|^2 + 2\gamma L_{gxy}^2\delta_\epsilon^2 r_v^4, \end{aligned} \tag{67}$$

According to the line 3 of Algorithm 1, we have $u_t = \nabla_y g(x_t, y_t)$. Then by using Lemma A.5, we have

$$\begin{aligned}
&g(x_{t+1}, y_{t+1}) - G(x_{t+1}) - \big(g(x_t, y_t) - G(x_t)\big) \\
&\le -\frac{\lambda\mu}{2}\big(g(x_t, y_t) - G(x_t)\big) + \frac{1}{8\gamma}\|x_{t+1} - x_t\|^2 - \frac{1}{4\lambda}\|y_{t+1} - y_t\|^2 + \lambda\|\nabla_y g(x_t, y_t) - u_t\|^2 \\
&= -\frac{\lambda\mu}{2}\big(g(x_t, y_t) - G(x_t)\big) + \frac{\gamma}{8}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 - \frac{1}{4\lambda}\|y_{t+1} - y_t\|^2, \qquad (68)
\end{aligned}$$

where the above equality holds by $u_t = \nabla_y g(x_t, y_t)$ and $\mathcal{G}(x_t, w_t, \gamma) = \frac{1}{\gamma}(x_t - x_{t+1})$.

By using Lemma A.3, we have

$$\begin{aligned}
&\|v_{t+1} - v_{t+1}^*\|^2 - \|v_t - v_t^*\|^2 \\
&\le -\frac{\mu\tau}{4}\|v_t - v_t^*\|^2 - \frac{3}{4}\|v_{t+1} - v_t\|^2 + \frac{25\tau L_{gyy}^2 r_v^4 \delta_\epsilon^2}{6\mu} \\
&\quad + \frac{20}{3}\big(\frac{L_f^2}{\mu^2} + \frac{L_{gyy}^2 C_{fx}^2}{\mu^4}\big)\big(\|x_{t+1} - x_t\|^2 + \|y_{t+1} - y_t\|^2\big) \\
&= -\frac{\mu\tau}{4}\|v_t - v_t^*\|^2 - \frac{3}{4}\|v_{t+1} - v_t\|^2 + \frac{25\tau L_{gyy}^2 r_v^4 \delta_\epsilon^2}{6\mu} + \frac{20}{3}\breve{L}^2\gamma^2\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + \frac{20}{3}\breve{L}^2\|y_{t+1} - y_t\|^2, \qquad (69)
\end{aligned}$$

where the above equality holds by $\breve{L}^2 = \frac{L_f^2}{\mu^2} + \frac{L_{gyy}^2 C_{fx}^2}{\mu^4}$ and $\mathcal{G}(x_t, w_t, \gamma) = \frac{1}{\gamma}(x_t - x_{t+1})$.

Next we define a useful Lyapunov function (i.e. potential function) for any $t \ge 1$

$$\Psi_t = \Phi(x_t) + g(x_t, y_t) - G(x_t) + \|v_t - v_t^*\|^2. \qquad (70)$$

By using the above inequalities (67), (68) and (69), we have

$$\begin{aligned}
\Psi_{t+1} - \Psi_t &= \Phi(x_{t+1}) - \Phi(x_t) + g(x_{t+1}, y_{t+1}) - G(x_{t+1}) - \big(g(x_t, y_t) - G(x_t)\big) + \|v_{t+1} - v_{t+1}^*\|^2 - \|v_t - v_t^*\|^2 \\
&\le -\frac{\gamma}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + \frac{12\gamma}{\mu}\big(L_f^2 + r_v^2 L_{gxy}^2\big)\big(g(x_t, y_t) - G(x_t)\big) + 6\gamma C_{gxy}^2\|v_t - v_t^*\|^2 + 2\gamma L_{gxy}^2\delta_\epsilon^2 r_v^4 \\
&\quad - \frac{\lambda\mu}{2}\big(g(x_t, y_t) - G(x_t)\big) + \frac{\gamma}{8}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 - \frac{1}{4\lambda}\|y_{t+1} - y_t\|^2 \\
&\quad - \frac{\mu\tau}{4}\|v_t - v_t^*\|^2 - \frac{3}{4}\|v_{t+1} - v_t\|^2 + \frac{25\tau L_{gyy}^2 r_v^4 \delta_\epsilon^2}{6\mu} + \frac{20}{3}\breve{L}^2\gamma^2\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + \frac{20}{3}\breve{L}^2\|y_{t+1} - y_t\|^2 \\
&\le -\big(\frac{\lambda\mu}{2} - \frac{12\gamma}{\mu}\big(L_f^2 + r_v^2 L_{gxy}^2\big)\big)\big(g(x_t, y_t) - G(x_t)\big) - \frac{\gamma}{4}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 - \big(\frac{\mu\tau}{4} - 6\gamma C_{gxy}^2\big)\|v_t - v_t^*\|^2 \\
&\quad + 2\gamma L_{gxy}^2\delta_\epsilon^2 r_v^4 + \frac{25\tau L_{gyy}^2 r_v^4 \delta_\epsilon^2}{6\mu}, \qquad (71)
\end{aligned}$$

where the last inequality is due to $0 < \gamma \le \frac{3}{160\breve{L}^2}$ and $0 < \lambda \le \frac{3}{80\breve{L}^2}$.

Let

$$x_{t+1}^+ = \mathbb{P}_{\phi(\cdot)}^{\gamma}(x_t, \nabla F(x_t)) = \arg\min_{x\in\mathbb{R}^d}\Big\{\langle\nabla F(x_t), x\rangle + \frac{1}{2\gamma}\|x - x_t\|^2 + \phi(x)\Big\}, \qquad (72)$$

$$x_{t+1} = \mathbb{P}_{\phi(\cdot)}^{\gamma}(x_t, w_t) = \arg\min_{x\in\mathbb{R}^d}\Big\{\langle w_t, x\rangle + \frac{1}{2\gamma}\|x - x_t\|^2 + \phi(x)\Big\}. \qquad (73)$$

By the optimality conditions of (72) and (73), for any $z \in \mathbb{R}^d$, there exist $\vartheta_1 \in \partial\phi(x_{t+1}^+)$ and $\vartheta_2 \in \partial\phi(x_{t+1})$ such that

$$\langle\nabla F(x_t) + \frac{1}{\gamma}(x_{t+1}^+ - x_t) + \vartheta_1, z - x_{t+1}^+\rangle \ge 0, \qquad (74)$$

$$\langle w_t + \frac{1}{\gamma}(x_{t+1} - x_t) + \vartheta_2, z - x_{t+1}\rangle \ge 0. \qquad (75)$$

Putting $z = x_{t+1}$ into (74), by the convexity of $\phi(x)$, we have

$$\begin{aligned}\langle \nabla F(x_t), x_{t+1} - x_{t+1}^+ \rangle &\geq \frac{1}{\gamma}\langle x_{t+1}^+ - x_t, x_{t+1}^+ - x_{t+1}\rangle + \langle \vartheta_1, x_{t+1}^+ - x_{t+1}\rangle \\ &\geq \frac{1}{\gamma}\langle x_{t+1}^+ - x_t, x_{t+1}^+ - x_{t+1}\rangle + \phi(x_{t+1}^+) - \phi(x_{t+1}). \end{aligned} \tag{76}$$

Similarly, putting $z = x_{t+1}^+$ into (75), by the convexity of $\phi(x)$, we have

$$\begin{aligned}\langle w_t, x_{t+1}^+ - x_{t+1} \rangle &\geq \frac{1}{\gamma}\langle x_{t+1} - x_t, x_{t+1} - x_{t+1}^+\rangle + \langle \vartheta_2, x_{t+1} - x_{t+1}^+\rangle \\ &\geq \frac{1}{\gamma}\langle x_{t+1} - x_t, x_{t+1} - x_{t+1}^+\rangle + \phi(x_{t+1}) - \phi(x_{t+1}^+). \end{aligned} \tag{77}$$

Summing up (76) and (77), we can obtain

$$\|\nabla F(x_t) - w_t\|\|x_{t+1} - x_{t+1}^+\| \geq \langle \nabla F(x_t) - w_t, x_{t+1} - x_{t+1}^+\rangle \geq \frac{1}{\gamma}\|x_{t+1}^+ - x_{t+1}\|^2. \tag{78}$$

Then we have

$$\|\nabla F(x_t) - w_t\| \geq \frac{1}{\gamma}\|x_{t+1}^+ - x_{t+1}\| = \frac{1}{\gamma}\left\|\mathbb{P}_{\phi(\cdot)}^{\gamma}(x_t, \nabla F(x_t)) - \mathbb{P}_{\phi(\cdot)}^{\gamma}(x_t, w_t)\right\|. \tag{79}$$

Since $\mathcal{G}(x_t, w_t, \gamma) = \frac{1}{\gamma}\big(x_t - \mathbb{P}_{\phi(\cdot)}^{\gamma}(x_t, w_t)\big)$ and $\mathcal{G}(x_t, \nabla F(x_t), \gamma) = \frac{1}{\gamma}\big(x_t - \mathbb{P}_{\phi(\cdot)}^{\gamma}(x_t, \nabla F(x_t))\big)$, we have

$$\begin{aligned}\|\mathcal{G}(x_t, \nabla F(x_t), \gamma)\|^2 &\leq 2\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + 2\|\mathcal{G}(x_t, w_t, \gamma) - \mathcal{G}(x_t, \nabla F(x_t), \gamma)\|^2 \\ &= 2\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + \frac{2}{\gamma^2}\|\mathbb{P}_{\phi(\cdot)}^{\gamma}(x_t, \nabla F(x_t)) - \mathbb{P}_{\phi(\cdot)}^{\gamma}(x_t, w_t)\|^2 \\ &\overset{(i)}{\leq} 2\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + 2\|w_t - \nabla F(x_t))\|^2 \\ &\leq 2\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + \frac{24}{\mu}\big(L_f^2 + r_v^2 L_{gxy}^2\big)\big(g(x_t, y_t) - G(x_t)\big) + 12C_{gxy}^2\|v_t - v_t^*\|^2 \\ &\quad + 4L_{gxy}^2\delta_\epsilon^2 r_v^4, \end{aligned} \tag{80}$$

where the inequality (i) holds by the above inequality (79). Then we can obtain

$$\begin{aligned}-\|\mathcal{G}(x_t, w_t, \gamma)\|^2 \leq -\frac{1}{2}\|\mathcal{G}(x_t, \nabla F(x_t), \gamma)\|^2 + \frac{12}{\mu}\big(L_f^2 + r_v^2 L_{gxy}^2\big)\big(g(x_t, y_t) - G(x_t)\big) + 6C_{gxy}^2\|v_t - v_t^*\|^2 \\ + 2L_{gxy}^2\delta_\epsilon^2 r_v^4. \end{aligned} \tag{81}$$

Plugging the above inequalities (81) into (71), we can further get

$$\begin{aligned}
&\Psi_{t+1}-\Psi_t\\
&\leq -\big(\frac{\lambda\mu}{2}-\frac{12\gamma}{\mu}\big(L_f^2+r_v^2L_{gxy}^2\big)\big)\big(g(x_t,y_t)-G(x_t)\big)-\frac{\gamma}{4}\|\mathcal{G}(x_t,w_t,\gamma)\|^2-\big(\frac{\mu\tau}{4}-6\gamma C_{gxy}^2\big)\|v_t-v_t^*\|^2\\
&\quad+2\gamma L_{gxy}^2\delta_\epsilon^2r_v^4+\frac{25\tau L_{gyy}^2r_v^4\delta_\epsilon^2}{6\mu}\\
&\leq -\big(\frac{\lambda\mu}{2}-\frac{12\gamma}{\mu}\big(L_f^2+r_v^2L_{gxy}^2\big)\big)\big(g(x_t,y_t)-G(x_t)\big)-\frac{\gamma}{8}\|\mathcal{G}(x_t,\nabla F(x_t),\gamma)\|^2\\
&\quad+\frac{3\gamma}{\mu}\big(L_f^2+r_v^2L_{gxy}^2\big)\big(g(x_t,y_t)-G(x_t)\big)+\frac{3\gamma C_{gxy}^2}{2}\|v_t-v_t^*\|^2+\frac{\gamma L_{gxy}^2\delta_\epsilon^2r_v^4}{2}\\
&\quad-\big(\frac{\mu\tau}{4}-6\gamma C_{gxy}^2\big)\|v_t-v_t^*\|^2+2\gamma L_{gxy}^2\delta_\epsilon^2r_v^4+\frac{25\tau L_{gyy}^2r_v^4\delta_\epsilon^2}{6\mu}\\
&=-\big(\frac{\lambda\mu}{2}-\frac{15\gamma}{\mu}\big(L_f^2+r_v^2L_{gxy}^2\big)\big)\big(g(x_t,y_t)-G(x_t)\big)-\frac{\gamma}{8}\|\mathcal{G}(x_t,\nabla F(x_t),\gamma)\|^2\\
&\quad-\big(\frac{\mu\tau}{4}-\frac{15\gamma C_{gxy}^2}{2}\big)\|v_t-v_t^*\|^2+\frac{5\gamma L_{gxy}^2\delta_\epsilon^2r_v^4}{2}+\frac{25\tau L_{gyy}^2r_v^4\delta_\epsilon^2}{6\mu}\\
&\leq -\frac{\gamma}{8}\|\mathcal{G}(x_t,\nabla F(x_t),\gamma)\|^2+\frac{5\gamma L_{gxy}^2\delta_\epsilon^2r_v^4}{2}+\frac{25\tau L_{gyy}^2r_v^4\delta_\epsilon^2}{6\mu},
\end{aligned}\tag{82}$$

where the last inequality holds by $0<\gamma\leq\big(\frac{\mu\tau}{30C_{gxy}^2},\frac{\mu^2\lambda}{30(L_f^2+r_v^2L_{gxy}^2)}\big)$.

Based on the inequality (82), we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T\|\mathcal{G}(x_t,\nabla F(x_t),\gamma)\|^2&\leq\frac{1}{T}\sum_{t=1}^T\frac{8(\Psi_t-\Psi_{t+1})}{\gamma}+20L_{gxy}^2\delta_\epsilon^2r_v^4+\frac{100\tau L_{gyy}^2r_v^4\delta_\epsilon^2}{3\gamma\mu}\\
&\overset{(i)}{\leq}\frac{8(\Psi_1-\Phi^*)}{T\gamma}+20L_{gxy}^2\delta_\epsilon^2r_v^4+\frac{100\tau L_{gyy}^2r_v^4\delta_\epsilon^2}{3\gamma\mu}\\
&=\frac{8(\Phi(x_1)+g(x_1,y_1)-G(x_1)+\|v_1-v_1^*\|^2-\Phi^*)}{T\gamma}+20L_{gxy}^2\delta_\epsilon^2r_v^4\\
&\quad+\frac{100\tau L_{gyy}^2r_v^4\delta_\epsilon^2}{3\gamma\mu},
\end{aligned}\tag{83}$$

where the above inequality (i) holds by Assumption 2.5.

Set $\delta_\epsilon=O\big(\frac{1}{\sqrt{T}\max(L_{gxy}^2,L_{gyy}^2/\mu)r_v^2}\big)$, we can obtain

$$\min_{1\leq t\leq T}\|\mathcal{G}(x_t,\nabla F(x_t),\gamma)\|^2\leq\frac{1}{T}\sum_{t=1}^T\|\mathcal{G}(x_t,\nabla F(x_t),\gamma)\|^2\leq O(\frac{1}{T}).\tag{84}$$

□

A.2. Convergence Analysis of of HJFBiO Algorithm for Bilevel Optimization without Regularization

In this subsection, we provide the convergence analysis of our HJFBiO algorithm for bilevel optimization without Regularization.

Lemma A.7. *Assume the sequence $\{x_t,y_t,v_t\}_{t=1}^T$ be generated from Algorithm 1, given $0<\gamma\leq\frac{1}{2L_F}$, we have*

$$\begin{aligned}
F(x_{t+1})\leq F(x_t)-\frac{\gamma}{2}\|\nabla F(x_t)\|^2-\frac{\gamma}{4}\|\widetilde{\nabla}f(x_t,y_t,v_t)\|^2+\gamma L_{gxy}^2r_v^4\delta_\epsilon^2\\
+\frac{6\gamma}{\mu}\big(L_f^2+r_v^2L_{gxy}^2\big)\big(g(x_t,y_t)-G(x_t)\big)+3\gamma C_{gxy}^2\|v_t-v_t^*\|^2,
\end{aligned}$$

Proof. When $\phi(x) \equiv 0$, at the line 4 of Algorithm 1, we have $x_{t+1} = x_t - \gamma w_t$. By using the Lipschitz smoothness of the objective function $F(x)$, we have

$$\begin{aligned}
F(x_{t+1}) &\leq F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t\rangle + \frac{L_F}{2}\|x_{t+1} - x_t\|^2 \\
&= F(x_t) - \gamma\langle \nabla F(x_t), w_t\rangle + \frac{\gamma^2 L_F}{2}\|w_t\|^2 \\
&= F(x_t) - \frac{\gamma}{2}\|\nabla F(x_t)\|^2 - \frac{\gamma}{2}(1-\gamma L_F)\|w_t\|^2 + \frac{\gamma}{2}\|w_t - \nabla F(x_t)\|^2 \\
&= F(x_t) - \frac{\gamma}{2}\|\nabla F(x_t)\|^2 - \frac{\gamma}{2}(1-\gamma L_F)\|\widetilde{\nabla} f(x_t, y_t, v_t)\|^2 + \frac{\gamma}{2}\|\widetilde{\nabla} f(x_t, y_t, v_t) - \nabla F(x_t)\|^2 \\
&\overset{(i)}{\leq} F(x_t) - \frac{\gamma}{2}\|\nabla F(x_t)\|^2 - \frac{\gamma}{4}\|\widetilde{\nabla} f(x_t, y_t, v_t)\|^2 + \frac{\gamma}{2}\|\widetilde{\nabla} f(x_t, y_t, v_t) - \nabla F(x_t)\|^2 \\
&\overset{(ii)}{\leq} F(x_t) - \frac{\gamma}{2}\|\nabla F(x_t)\|^2 - \frac{\gamma}{4}\|\widetilde{\nabla} f(x_t, y_t, v_t)\|^2 + \gamma L_{gxy}^2 r_v^4 \delta_\epsilon^2 \\
&\quad + \frac{6\gamma}{\mu}\big(L_f^2 + r_v^2 L_{gxy}^2\big)\big(g(x_t, y_t) - G(x_t)\big) + 3\gamma C_{gxy}^2\|v_t - v_t^*\|^2,
\end{aligned} \tag{85}$$

where the above inequality (i) holds by $0 < \gamma \leq \frac{1}{2L_F}$, and the above inequality (ii) holds by the above inequality (52).

□

Theorem A.8. *Assume the sequence $\{x_t, y_t, v_t\}_{t=1}^T$ be generated from our Algorithm 1. Under the above Assumptions 2.1-2.5, let $0 < \gamma \leq \min\Big(\frac{1}{2L_F}, \frac{\lambda\mu}{16L_G}, \frac{\mu}{16L_g^2}, \frac{3}{160\breve{L}^2}, \frac{\mu\tau}{12C_{gxy}^2}, \frac{\lambda\mu^2}{12\big(L_f^2 + r_v^2 L_{gxy}^2\big)}\Big)$, $0 < \lambda \leq \min\big(\frac{1}{2L_g}, \frac{3}{80\breve{L}^2}\big)$ and $0 < \tau \leq \frac{1}{6L_g}$, we have*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T \|\nabla F(x_t)\|^2 \\
&\leq \frac{2\big(F(x_1) + g(x_1, y_1) - G(x_1) + \|v_1 - v_1^*\|^2 - F^*\big)}{T\gamma} + 2L_{gxy}^2 r_v^4 \delta_\epsilon^2 + \frac{25\tau L_{gyy}^2 r_v^4 \delta_\epsilon^2}{2\gamma\mu},
\end{aligned} \tag{86}$$

where $\breve{L}^2 = \frac{L_f^2}{\mu^2} + \frac{L_{gyy}^2 C_{fx}^2}{\mu^4}$ and $F^* = \inf_{x\in\mathbb{R}^d} F(x) > -\infty$.

Proof. According to the line 3 of Algorithm 1, we have $u_t = \nabla_y g(x_t, y_t)$. Then by using Lemma A.5, we have

$$\begin{aligned}
&g(x_{t+1}, y_{t+1}) - G(x_{t+1}) - \big(g(x_t, y_t) - G(x_t)\big) \\
&\leq -\frac{\lambda\mu}{2}\big(g(x_t, y_t) - G(x_t)\big) + \frac{1}{8\gamma}\|x_{t+1} - x_t\|^2 - \frac{1}{4\lambda}\|y_{t+1} - y_t\|^2 + \lambda\|\nabla_y g(x_t, y_t) - u_t\|^2 \\
&= -\frac{\lambda\mu}{2}\big(g(x_t, y_t) - G(x_t)\big) + \frac{\gamma}{8}\|\widetilde{\nabla} f(x_t, y_t, v_t)\|^2 - \frac{1}{4\lambda}\|y_{t+1} - y_t\|^2,
\end{aligned} \tag{87}$$

where the above equality holds by $u_t = \nabla_y g(x_t, y_t)$ and $x_{t+1} = x_t - \gamma w_t = x_t - \gamma\widetilde{\nabla} f(x_t, y_t, v_t)$.

By using Lemma A.3, we have

$$\begin{aligned}
&\|v_{t+1} - v_{t+1}^*\|^2 - \|v_t - v_t^*\|^2 \\
&\leq -\frac{\mu\tau}{4}\|v_t - v_t^*\|^2 - \frac{3}{4}\|v_{t+1} - v_t\|^2 + \frac{25\tau L_{gyy}^2 r_v^4 \delta_\epsilon^2}{6\mu} \\
&\quad + \frac{20}{3}\big(\frac{L_f^2}{\mu^2} + \frac{L_{gyy}^2 C_{fx}^2}{\mu^4}\big)\big(\|x_{t+1} - x_t\|^2 + \|y_{t+1} - y_t\|^2\big) \\
&= -\frac{\mu\tau}{4}\|v_t - v_t^*\|^2 - \frac{3}{4}\|v_{t+1} - v_t\|^2 + \frac{25\tau L_{gyy}^2 r_v^4 \delta_\epsilon^2}{6\mu} + \frac{20}{3}\breve{L}^2\gamma^2\|\widetilde{\nabla} f(x_t, y_t, v_t)\|^2 + \frac{20}{3}\breve{L}^2\|y_{t+1} - y_t\|^2,
\end{aligned} \tag{88}$$

where the above equality holds by $x_{t+1} = x_t - \gamma w_t = x_t - \gamma\widetilde{\nabla} f(x_t, y_t, v_t)$ and $\breve{L}^2 = \frac{L_f^2}{\mu^2} + \frac{L_{gyy}^2 C_{fx}^2}{\mu^4}$. Then by using Lemma A.7, we have

$$F(x_{t+1}) - F(x_t) \leq -\frac{\gamma}{2}\|\nabla F(x_t)\|^2 - \frac{\gamma}{4}\|\widetilde{\nabla} f(x_t, y_t, v_t)\|^2 + \gamma L_{gxy}^2 r_v^4 \delta_\epsilon^2 + \frac{6\gamma}{\mu}\big(L_f^2 + r_v^2 L_{gxy}^2\big)\big(g(x_t, y_t) - G(x_t)\big) + 3\gamma C_{gxy}^2 \|v_t - v_t^*\|^2. \tag{89}$$

Next, we define a useful Lyapunov function (i.e. potential function), for any $t \geq 1$

$$\Omega_t = F(x_t) + g(x_t, y_t) - G(x_t) + \|v_t - v_t^*\|^2. \tag{90}$$

By combining the above inequalities (87), (88) and (89), we have

$$\begin{aligned}
&\Omega_{t+1} - \Omega_t \\
&= F(x_{t+1}) - F(x_t) + g(x_{t+1}, y_{t+1}) - G(x_{t+1}) - \big(g(x_t, y_t) - G(x_t)\big) + \|v_{t+1} - v_{t+1}^*\|^2 - \|v_t - v_t^*\|^2 \\
&\leq -\frac{\gamma}{2}\|\nabla F(x_t)\|^2 - \frac{\gamma}{4}\|\widetilde{\nabla} f(x_t, y_t, v_t)\|^2 + \gamma L_{gxy}^2 r_v^4 \delta_\epsilon^2 + \frac{6\gamma}{\mu}\big(L_f^2 + r_v^2 L_{gxy}^2\big)\big(g(x_t, y_t) - G(x_t)\big) \\
&\quad + 3\gamma C_{gxy}^2 \|v_t - v_t^*\|^2 - \frac{\lambda\mu}{2}\big(g(x_t, y_t) - G(x_t)\big) + \frac{\gamma}{8}\|\widetilde{\nabla} f(x_t, y_t, v_t)\|^2 - \frac{1}{4\lambda}\|y_{t+1} - y_t\|^2 \\
&\quad - \frac{\mu\tau}{4}\|v_t - v_t^*\|^2 - \frac{3}{4}\|v_{t+1} - v_t\|^2 + \frac{25\tau L_{gyy}^2 r_v^4 \delta_\epsilon^2}{6\mu} + \frac{20}{3}\breve{L}^2\gamma^2\|\widetilde{\nabla} f(x_t, y_t, v_t)\|^2 + \frac{20}{3}\breve{L}^2\|y_{t+1} - y_t\|^2 \\
&\leq -\frac{\gamma}{2}\|\nabla F(x_t)\|^2 - \big(\frac{\gamma}{8} - \frac{20}{3}\breve{L}^2\gamma^2\big)\|\widetilde{\nabla} f(x_t, y_t, v_t)\|^2 - \Big(\frac{\lambda\mu}{2} - \frac{6\gamma}{\mu}\big(L_f^2 + r_v^2 L_{gxy}^2\big)\Big)\big(g(x_t, y_t) - G(x_t)\big) \\
&\quad - \big(\frac{1}{4\lambda} - \frac{20}{3}\breve{L}^2\big)\|y_{t+1} - y_t\|^2 - \big(\frac{\mu\tau}{4} - 3\gamma C_{gxy}^2\big)\|v_t - v_t^*\|^2 + \gamma L_{gxy}^2 r_v^4 \delta_\epsilon^2 + \frac{25\tau L_{gyy}^2 r_v^4 \delta_\epsilon^2}{6\mu} \\
&\leq -\frac{\gamma}{2}\|\nabla F(x_t)\|^2 + \gamma L_{gxy}^2 r_v^4 \delta_\epsilon^2 + \frac{25\tau L_{gyy}^2 r_v^4 \delta_\epsilon^2}{6\mu},
\end{aligned} \tag{91}$$

where the last inequality holds by $0 < \gamma \leq \min\Big(\frac{3}{160\breve{L}^2}, \frac{\mu\tau}{12C_{gxy}^2}, \frac{\lambda\mu^2}{12\left(L_f^2 + r_v^2 L_{gxy}^2\right)}\Big)$ and $0 < \lambda \leq \frac{3}{80\breve{L}^2}$. Then we can obtain

$$\|\nabla F(x_t)\|^2 \leq \frac{2\big(\Omega_t - \Omega_{t+1}\big)}{\gamma} + 2L_{gxy}^2 r_v^4 \delta_\epsilon^2 + \frac{25\tau L_{gyy}^2 r_v^4 \delta_\epsilon^2}{2\gamma\mu}. \tag{92}$$

Averaging the above inequality (92), we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T \|\nabla F(x_t)\|^2 &\leq \frac{2\big(\Omega_1 - \Omega_{T+1}\big)}{T\gamma} + 2L_{gxy}^2 r_v^4 \delta_\epsilon^2 + \frac{25\tau L_{gyy}^2 r_v^4 \delta_\epsilon^2}{2\gamma\mu} \\
&\overset{(i)}{\leq} \frac{2(\Omega_1 - F^*)}{T\gamma} + 2L_{gxy}^2 r_v^4 \delta_\epsilon^2 + \frac{25\tau L_{gyy}^2 r_v^4 \delta_\epsilon^2}{2\gamma\mu} \\
&= \frac{2\big(F(x_1) + g(x_1, y_1) - G(x_1) + \|v_1 - v_1^*\|^2 - F^*\big)}{T\gamma} + 2L_{gxy}^2 r_v^4 \delta_\epsilon^2 + \frac{25\tau L_{gyy}^2 r_v^4 \delta_\epsilon^2}{2\gamma\mu},
\end{aligned} \tag{93}$$

where the above inequality (i) holds by $F^* = \inf_{x\in\mathbb{R}^d} F(x) > -\infty$.

Set $\delta_\epsilon = O\big(\frac{1}{\sqrt{T}\max(L_{gxy}^2, L_{gyy}^2/\mu) r_v^2}\big)$, we can obtain

$$\min_{1\leq t\leq T}\|\nabla F(x_t)\|^2 \leq \frac{1}{T}\sum_{t=1}^T \|\nabla F(x_t)\|^2 \leq O(\frac{1}{T}). \tag{94}$$

□

B. Related Works

The GALET method (Xiao et al., 2023) is meaningless for nonconvex-PL bilevel optimization, which is based on the following facts:

1) In the convergence analysis, the GALET method simultaneously uses the PL condition, its Assumption 2 (i.e., let $\sigma_g = \inf_{x,y}\{\sigma^+_{\min}(\nabla^2_{yy}g(x,y))\} > 0$ for all (x,y)) and its Assumption 1 (i.e., $\nabla^2_{yy}g(x,y)$ is Lipschitz continuous).
2) In the nonconvex-PL bilevel optimization problems, Hessian matrix $\nabla^2_{yy}g(x,y))$ has two cases: **the first case**: $\nabla^2_{yy}g(x,y)$ is singular; **the second case**: $\nabla^2_{yy}g(x,y)$ is not singular.
3) **The first case**: $\nabla^2_{yy}g(x,y)$ is singular. Since $\nabla^2_{yy}g(x,y)$ is Lipschitz continuous by Assumption 1 of (Xiao et al., 2023), the singular-value of $\nabla^2_{yy}g(x,y)$ also is continuous. Thus, combining its Assumption 1 with Assumption 2 imply that the lower bound of the non-zero singular values $\sigma_g = \inf_{x,y}\{\sigma^+_{\min}(\nabla^2_{yy}g(x,y))\} > 0$ is close to zero. Under this case, the constant $L_w = \frac{\ell_{f,1}}{\sigma_g} + \frac{\sqrt{2}\ell_{g,2}\ell_{f,0}}{\sigma_g^2} \to +\infty$ used in its Lemmas 6 and 9, and $L_F = \ell_{f,0}(\ell_{f,1}+\ell_{g,2})/\sigma_g \to +\infty$ used in its Lemma 12.
4) **The second case**: $\nabla^2_{yy}g(x,y)$ is not singular. By Assumption 2 of (Xiao et al., 2023), the singular values of Hessian is bounded away from 0, i.e., $\sigma_g > 0$. Under the this case, the PL condition, Lipschitz continuous of Hessian and its Assumption 2 (the singular values of Hessian is bounded away from 0, i.e., $\sigma_g > 0$) imply that GALET uses strongly convex assumption on $g(x,y)$ at variable y. **Note that** although the singular values of Hessian $\nabla^2_{yy}g(x,y)$ exclude zero, i.e, the eigenvalues of Hessian $\nabla^2_{yy}g(x,y)$ may be in $[-\ell_{g,2},-\sigma_g]\bigcup[\sigma_g,\ell_{g,2}]$, we cannot have negative eigenvalues at the minimizer $y^*(x)$. Meanwhile, since Hessian is Lipschitz continuous, its all eigenvalues are in $[\sigma_g,\ell_{g,2}]$. Thus, the PL condition, Lipschitz continuous of Hessian and its Assumption 2 (the singular values of Hessian is bounded away from 0, i.e., $\sigma_g > 0$) imply that the GALET assumes the strongly convex.

Although (Kwon et al., 2023) studied the nonconvex-PL bilevel optimization, it also requires some relatively strict assumptions (e.g., Assumption 1, 4,5,6,7,8 of (Kwon et al., 2023)). For example, Assumption 1 of (Kwon et al., 2023) gives proximal error-bound (EB) condition that is analogous to PL condition, and its Assumption 4 (2) requires the bounded $|f(x,y)|$. In particular, its Assumption 7 (2) assumes the upper-level function $f(x,y)$ has Lipschitz Hessian. Under these conditions, (Kwon et al., 2023) has a gradient complexity of $O(\epsilon^{-1.5})$ for finding an ϵ-stationary solution of nonconvex-PL bilevel optimization. However, without relying on Lipschitz Hessian of function $f(x,y)$ and bounded $|f(x,y)|$, our algorithm obtains an optimal gradient complexity of $O(\epsilon^{-1})$, which matches the lower bound established by the first-order method for finding an ϵ-stationary point of nonconvex smooth optimization (Carmon et al., 2020).

Meanwhile, (Chen et al., 2024) studied the nonconvex-PL bilevel optimization, but it also relies on some strict assumptions, e.g., $h_\sigma = \sigma f + g$ is μ-PL (Please see the Assumption 4.1 (a) of (Chen et al., 2024)). While our paper only assumes the lower-level function g is μ-PL. When $\sigma > 0$, Assumption 4.1 (a) of (Chen et al., 2024) is stricter than our assumption (the lower-level function g is μ-PL). In particular, the upper-level function $f(x,y)$ also requires Lipschitz Hessian (Please see the Assumption 4.1 (d) of (Chen et al., 2024)).

Note that from (Carmon et al., 2020), the optimal gradient complexity is $O(\epsilon^{-0.75})$ (or $O(\epsilon^{-1.5})$) for finding an ϵ-stationary point of smooth nonconvex optimization problem $\min_x f(x)$ with Lipschitz Hessian condition, i.e, $\|\nabla f(x)\|^2 \leq \epsilon$ (or $\|\nabla f(x)\| \leq \epsilon$). Meanwhile, based on Lipschitz Hessian of function $f(x,y)$, (Yang et al., 2023a) can obtain a lower gradient complexity of $O(\epsilon^{-0.875})$ (or $O(\epsilon^{-1.75})$) for finding an ϵ-stationary point of nonconvex strongly-convex bilevel optimization, i.e., $\|\nabla\Phi(x)\|^2 \leq \epsilon$ (or $\|\nabla\Phi(x)\| \leq \epsilon$). Under these strict assumptions, thus, although (Chen et al., 2024) also obtain a gradient complexity of $O(\epsilon^{-1})$, *this is not optimal gradient complexity.*

Lemma B.1. *(Lemma G.6 of (Chen et al., 2024)) For a μ-PL function $h(x):\mathbb{R}^d \to \mathbb{R}$ that is twice differentiable, at any $x^* \in \arg\min_x h(x)$,*

$$\lambda^+_{\min}(\nabla^2 h(x^*)) \geq \mu, \tag{95}$$

where $\lambda^+_{\min}(\cdot)$ denotes the smallest non-zero eigenvalue.

In fact, Lemma G.6 of (Chen et al., 2024) is exactly useful for our HJFBiO method. Based on Lemma G.6 of (Chen et al., 2024), our Assumption 2.2 is reasonable when has an unique minimizer. Meanwhile, our Assumption 4.8 also is reasonable when have multiple local minimizers.