Double Momentum Method for Lower-Level Constrained Bilevel Optimization

Wanli Shi [1 2] **Yi Chang** [1] **Bin Gu** [1 2]

Abstract

Bilevel optimization (BO) has recently gained prominence in many machine learning applications due to its ability to capture the nested structure inherent in these problems. Recently, many hypergradient methods have been proposed as effective solutions for solving large-scale problems. However, current hypergradient methods for the lower-level constrained bilevel optimization (LCBO) problems need very restrictive assumptions, namely, where optimality conditions satisfy the differentiability and invertibility conditions and lack a solid analysis of the convergence rate. What's worse, existing methods require either double-loop updates, which are sometimes less efficient. To solve this problem, in this paper, we propose a new hypergradient of LCBO leveraging the theory of nonsmooth implicit function theorem instead of using the restrive assumptions. In addition, we propose a *single-loop single-timescale* algorithm based on the double-momentum method and adaptive step size method and prove it can return a (δ, ϵ)-stationary point with $\tilde{\mathcal{O}}(d_2^2\epsilon^{-4})$ iterations. Experiments on two applications demonstrate the effectiveness of our proposed method.

1. Introduction

Bilevel optimization (BO) (Bard, 2013; Colson et al., 2007) plays a central role in various significant machine learning applications, including hyper-parameter optimization (Pedregosa, 2016; Bergstra et al., 2011; Bertsekas, 1976; Shi & Gu, 2021), meta-learning (Feurer et al., 2015; Franceschi et al., 2018; Rajeswaran et al., 2019), reinforcement learning (Hong et al., 2020; Konda & Tsitsiklis, 2000). Generally speaking, the BO can be formulated as follows,

$$\min_{x\in\mathcal{X}} F(x) = f(x, y^*(x)) \quad (1)$$
$$s.t.\ y^*(x) = \arg\min_{y\in\mathcal{Y}} g(x, y),$$

where $\mathcal{X}$ and $\mathcal{Y}$ are convex subsets in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. It involves a competition between two parties or two objectives, and if one party makes its choice first, it will affect the optimal choice of the other party.

Recently, hypergradient methods have shown great effectiveness in solving various large-scale bilevel optimization problems, where there is no constraint in the lower-level objective, *i.e.*, $\mathcal{Y} = \mathbb{R}^{d_2}$. Specifically, (Franceschi et al., 2017; Pedregosa, 2016; Ji et al., 2021) proposed several double-loop algorithms to solve the BO problems. They first apply the gradient methods to approximate the solution to the lower-level problem and then implicit differentiable methods (Pedregosa, 2016; Ji et al., 2021) or explicit differentiable methods (Franceschi et al., 2017) can be used to approximate the gradient of the upper-level objective w.r.t x, namely hypergradient, to update x. However, in some real-world applications, such as in a sequential game, the problems must be updated at the same time (Hong et al., 2020), which makes these methods unsuitable. To solve this problem, (Hong et al., 2020) propose a single-loop two-timescale method, which updates y and x alternately with stepsize η_y and η_x, respectively, designed with different timescales as $\lim_{k\to\infty} \eta_x/\eta_y = 0$. However, due to the nature of two-timescale updates, it incurs the sub-optimal complexity $\mathcal{O}(\epsilon^{-5})$ (Chen et al., 2021). To further improve the efficiency, (Huang & Huang, 2021; Khanduri et al., 2021; Chen et al., 2021; Guo et al., 2021) proposed single-loop single-timescale methods, where η_x/η_y is a constant. These methods have the complexity of $\tilde{\mathcal{O}}(\epsilon^{-4})$ ($\tilde{\mathcal{O}}$ means omitting logarithmic factors) or better ($\tilde{\mathcal{O}}(\epsilon^{-3})$) to achieve the stationary point. However, all these methods are limited to the bilevel optimization problem with unconstrained lower-level problems and require the upper-level objective function to be differentiable. They cannot be directly applied when constraints are present in the lower-level optimization, i.e., $\mathcal{Y} \neq \mathbb{R}^{d_2}$, as the upper-level objective function is naturally non-differentiable (Xu & Zhu, 2023).

To solve the lower-level constrained bilevel optimization problem, recently, several methods have been proposed to

[*]Equal contribution [1]School of Artificial Intelligence, Jilin University, China [2]Mohamed bin Zayed University of Artificial Intelligence, UAE. Correspondence to: Bin Gu <jsgubin@gmail.com>.

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

Table 1: Several representative hypergradient approximation methods for the lower-level constrained BO problem. (The last column shows iteration numbers to find a stationary point. The gray color is used to highlight the main limitations of the listed algorithms)

Method	$F(x)$	Loop	Timescale	LL. Constraint	Restrictive Conditions	Iterations
AiPOD (Xiao et al., 2023)	Smooth	Double	$\times$	Affine sets	Not need	$\tilde{\mathcal{O}}(\epsilon^{-4})$
IG-AL (Tsaknakis et al., 2022)	Nonsmooth	Double	$\times$	Half space	Not need	$\times$
IAPTT-GM (Liu et al., 2021)	Nonsmooth	Double	$\times$	Convex set	$y^*(x)$ is differentiable	$\times$
RMD-PCD (Bertrand et al., 2022)	Nonsmooth	Double	$\times$	Norm set	$y^*(x)$ is differentiable	$\times$
JaxOpt(Blondel et al., 2022)	Nonsmooth	Double	$\times$	Convex set	$y^*(x)$ is differentiable	$\times$
DMLCBO (Ours)	Nonsmooth	Single	Single	Convex set	Not need	$\tilde{\mathcal{O}}(d_2^2\epsilon^{-4})$

approximate the hypergradient, as shown in Table 1. Specifically, (Xiao et al., 2023) reformulate the affine-constrained lower-level problem into an unconstrained problem, and then solve the new lower-level problem and use the implicit differentiable method to approximate the hyper-gradient. However, their convergence analysis only focuses on the affine-constrained problem and cannot be extended to a more general case. (Tsaknakis et al., 2022) solve the inner problem with projection gradient and use the implicit differentiable method to approximate the hyper-gradient for the half-space-constrained BO problem. However, they only give the asymptotic convergence analysis for this special case. Since many methods of calculating the Jacobian of the projection operators have been proposed (Martins & Astudillo, 2016; Djolonga & Krause, 2017; Blondel et al., 2020; Niculae & Blondel, 2017; Vaiter et al., 2013; Cherkaoui et al., 2020), the explicit or implicit methods can also be used to approximate the hypergradient in the LCBO problems, such as (Liu et al., 2021; Bertrand et al., 2020; 2022; Blondel et al., 2022). However, all these methods are based on restrictive assumptions, namely, where optimality conditions satisfy the differentiability and invertibility conditions, and lack a solid analysis of the convergence rate. What's worse, these methods can not be utilized to solve the sequential game which is mentioned above. Therefore, it is still an open challenge to design a *single-loop single-timescale* method with convergence rate analysis for the lower-level constrained bilevel optimization problems.

To overcome these problems, we propose a novel *single-loop single-timescale* method with a convergence guarantee for the lower-level constrained BO problems. Specifically, instead of using the restive assumptions used in (Blondel et al., 2022; Bertrand et al., 2020; 2022), we leverage the theory of nonsmooth implicit function theorems (Clarke, 1990; Bolte et al., 2021) to propose a new hypergradient of LCBO. Then, we use the randomized smoothing and Neumann series to further approximate the hypergradient. Using this new hypergradient approximation, we propose a *single-loop single-timescale* algorithm based on the double-momentum method and adaptive step size method to update the lower- and upper-level variables simultaneously. Theoretically, we prove our methods can return a (δ, ϵ)-stationary point with $\tilde{\mathcal{O}}(d_2^2\epsilon^{-4})$ iterations. The experimental results in two applications demonstrate the effectiveness of our proposed method.

We summarized our contributions as follows.

1. Leveraging the theory of nonsmooth implicit function theorems, we propose a new method to calculate the hypergradient of LCBO without using restrictive assumptions.

2. We propose a new method to approximate the hypergradient based on randomized smoothing and the Neumann series. Using this hypergradient approximation, we propose a *single-loop single-timescale* algorithm for the lower-level constrained BO problems.

3. Existing hypergradient-based methods for solving the lower-level constrained BO problems usually lack theoretical analysis on convergence rate. We prove our methods can return a (δ, ϵ)-stationary point with $\tilde{\mathcal{O}}(d_2^2\epsilon^{-4})$ iterations.

4. We compare our method with several state-of-the-art methods for the lower-level constrained BO problems on two applications. The experimental results demonstrate the effectiveness of our proposed method.

2. Preliminaries

Notations. Here, we give several important notations used in this paper. $\|\cdot\|$ denotes the ℓ_2 norm for vectors and spectral norm for matrices. I_d denotes a d-dimensional identity matrix. $A^\top$ denotes transpose of matrix A. Given a convex set $\mathcal{X}$, we define a projection operation to $\mathcal{X}$ as $\mathcal{P}_{\mathcal{X}}(x') = \arg\min_{x\in\mathcal{X}} 1/2\|x - x'\|^2$. $co\mathcal{X}$ denotes the convex hull of the set $\mathcal{X}$.

2.1. Lower-level Constrained Bilevel Optimization

In this paper, we consider the following BO problems where the lower-level problem has convex constraints,

$$\min_{x\in\mathbb{R}^{d_1}} F(x) = f(x, y^*(x)) \quad (2)$$
$$s.t. \quad y^*(x) = \arg\min_{y\in\mathcal{Y}\subseteq\mathbb{R}^{d_2}} g(x, y).$$

Then, we introduce several mild assumptions on the Problem (2).

Assumption 2.1. The upper-level function $f(x, y)$ satisfies the following conditions:

1. $\nabla_x f(x, y)$ is L_f-Lipschitz continuous w.r.t. $(x, y) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ and $\nabla_y f(x, y)$ is L_f-Lipschitz continuous w.r.t $(x, y) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, where $L_f > 0$.
2. For any $x \in \mathbb{R}^{d_1}$ and $y \in \mathbb{R}^{d_2}$, we have $\|\nabla_y f(x, y)\| \leq C_{fy}$.

Assumption 2.2. The lower-level function $g(x, y)$ satisfies the following conditions:

1. For any $(x, y) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, $g(x, y)$ is twice continuously differentiable in (x, y).
2. Fix x, $\nabla_y g(x, y)$ is L_g-Lipschitz continuous w.r.t y for some $L_g > 0$.
3. Fix x, for any y, $g(x, y)$ is μ_g-strongly-convex in y for some $\mu_g > 0$.
4. $\nabla^2_{xy} g(x, y)$ is L_{gxy}-Lipschitz continuous w.r.t $(x, y) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ and $\nabla^2_{yy} g(x, y)$ is L_{gyy}-Lipschitz continuous w.r.t $(x, y) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, where $L_{gxy} > 0$ and $L_{gyy} > 0$.
5. For any $(x, y) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, we have $\|\nabla^2_{xy} g(x, y)\| \leq C_{gxy}$.

These assumptions are commonly used in bilevel optimization problems (Ghadimi & Wang, 2018; Hong et al., 2020; Ji et al., 2021; Chen et al., 2021; Khanduri et al., 2021; Guo et al., 2021).

2.2. Review of Unconstrained Bilevel Optimization Methods

For the upper-level objective, we can naturally derive the following gradient w.r.t x using the chain rule (which is defined as hypergradient),

$$\nabla F(x) = \nabla_x f(x, y^*(x)) + (\nabla y^*(x))^\top \nabla_y f(x, y^*(x)).$$

The crucial problem of obtaining the hypergradient is calculating $\nabla y^*(x)$. If the lower-level problem is unconstrained, using the implicit differentiation method and the optimal condition $\nabla_y g(x, y^*(x)) = 0$, it is easy to show that for a given $x \in \mathbb{R}^{d_1}$, the following equation holds (Ghadimi & Wang, 2018; Hong et al., 2020; Ji et al., 2021; Chen et al., 2021; Khanduri et al., 2021)

$$\nabla y^*(x) = [\nabla^2_{yy} g(x, y^*(x))]^{-1} \nabla^2_{yx} g(x, y^*(x)). \quad (3)$$

Substituting $\nabla y^*(x)$ into $\nabla F(x)$, we can obtain the hypergradient. Then, we update x and y alternately using the gradient method.

2.3. Review of Lower-Level Constrained Bilevel Optimization Problem

For the constrained lower-level problem, one common method is to use the projection gradient method, which has the following optimal condition,

$$y^*(x) = \mathcal{P}_{\mathcal{Y}}(y^*(x) - \eta \nabla_y g(x, y^*(x))), \quad (4)$$

where $\eta > 0$ denotes the step-size. Recently, (Blondel et al., 2022; Bertrand et al., 2020; 2022) assume $y^*(x)$ and $\mathcal{P}_{\mathcal{Y}}(\cdot)$ to be differentiable at some special points and use the reverse method or implicit gradient method to derive the hypergradient. However, these assumptions are relatively strong and highly limit the usage of these methods.

3. Proposed Method

In this section, we propose a new method to approximate the hypergradient using randomized smoothing that makes convergence analysis possible. Then, equipped with this hypergradient, we propose our *single-loop single-timescale* method to find a stationary point of the lower-level constrained bilevel problem.

3.1. Hypergrdient of Lower-level Constrained Bilevel Optimization Problem

Before we present our method to calculate the hypergradient, we give the following definition of generalized Jacobian and gradient (Clarke, 1990), which is important in our method.

Definition 3.1. Given a scaler-valued function $F : \mathbb{R}^n \to \mathbb{R}$, which is Lipschitz near a given point x of interest. The generalized gradient is defined as:

$$\partial F(x) = co\{\lim \nabla F(x_i) : x_i \to x, x_i \notin S, x_i \notin \Omega_F\},$$

where Ω_F is the set of points at which F fails to be differentiable and S is any other set of measure zero.

Definition 3.2. Given a vector-valued function $F : \mathbb{R}^n \to \mathbb{R}^m$, which is Lipschitz near a given point x of interest. Denote the set of points at which F fails to be differentiable by Ω_F and the usual $m \times n$ Jacobian matrix of partial derivatives by $JF(y)$ whenever y is a point at which the

necessary partial derivatives exist. The generalized Jacobian of F at point x, denoted as $\partial F(x)$, is the convex hull of all $m \times n$ matrices obtained as the limit of a sequence of the form $JF(x_i)$, where $x_i \to x$ and $x_i \notin \Omega_F$. Symbolically, then, one has

$$\partial F(x) = co\{\lim JF(x_i) : x_i \to x, x_i \notin \Omega_F\}.$$

According to (Clarke, 1990), if we want to use the Jacobian Chain Rule to calculate the generalized gradient $\partial F(x)$ of the LCBO, the most important thing is to ensure that $y^*(x)$ is Lipschitz continuous. Fortunately, under the strongly convex assumption of the lower-level problem in Assumptions 2.2, we can obtain the following result.

Lemma 3.3. *Under Assumptions 2.2, we have the optimal solution to the lower-level problem is Lipschitz continuous with constant L_g/μ_g.*

Then, using the Jacobian Chain Rule in (Clarke, 1990) and Lemma 3.3, we have

$$\partial F(x) = \nabla_x f(x, y^*(x)) + (\partial y^*(x))^\top \nabla_y f(x, y^*(x)). \quad (5)$$

To obtain the generalized gradient $\partial F(x)$, the most critical point is to calculate the generalized Jacobian $\partial y^*(x)$. Inspired by the previous methods (Blondel et al., 2022; Bertrand et al., 2020; 2022), we use the Corollary of the Jacobian Chain Rule in (Clarke, 1990) to calculate the generalized Jacobian on Eqn 4 on both sides, since both $y^*(x)$ and $\mathcal{P}_{\mathcal{Y}}(z^*)$ are Lipschitz continuous, where $z^* = y^*(x) - \eta \nabla_y g(x, y^*(x))$. Therefore, we can obtain

$$\partial y^*(x) v = \partial \mathcal{P}_{\mathcal{Y}}(z^*) \partial \left(y^*(x) - \eta \nabla_y g(x, y^*(x))\right) v. \quad (6)$$

where $v \in \mathbb{R}^{d_1}$. Using the same method, we have

$$\begin{aligned} &\partial \left(y^*(x) - \eta \nabla_y g(x, y^*(x))\right) v = \partial y^*(x) v \\ &- \eta(\nabla^2_{yx} g(x, y^*(x)) - \nabla^2_{yy} g(x, y^*(x)) \partial y^*(x)) v. \end{aligned} \quad (7)$$

Substituting Equation (7) into formula (6), we have

$$\begin{aligned} \partial y^*(x) v =& \partial \mathcal{P}_{\mathcal{Y}}(z^*)(\partial y^*(x) v - \eta(\nabla^2_{yx} g(x, y^*(x)) \\ &+ \nabla^2_{yy} g(x, y^*(x)) \partial y^*(x)) v). \end{aligned} \quad (8)$$

Suppose there exist $A \in \partial y^*(x)$ and $H \in \partial \mathcal{P}_{\mathcal{Y}}(z^*)$, which makes the following equality hold,

$$Av = H \cdot \left(A - \eta \nabla^2_{yx} g(x, y^*(x)) - \eta \nabla^2_{yy} g(x, y^*(x)) A\right) v. \quad (9)$$

Then, rearranging the above equality, we have

$$\begin{aligned} &\left[I_{d_2} - H \cdot \left(I_{d_2} - \eta \nabla^2_{yy} g(x, y^*(x))\right)\right] A \\ =& - \eta H \cdot \nabla^2_{yx} g(x, y^*(x)) \end{aligned} \quad (10)$$

To solve the above equation, the challenge is to ensure that $I_{d_2} - H \cdot \left(I_{d_2} - \eta \nabla^2_{yy} g(x, y^*(x))\right)$ is invertible. Since $\mathcal{P}_{\mathcal{Y}}(z^*)$ is non-expensive, we have $\|H\| \leq 1$. Since g is strongly convex, setting $\eta < 1/\mu_g$, we have $\|I_{d_2} - \eta \nabla^2_{yy} g(x, y^*(x))\| < 1$. Therefore, we have $\|H \cdot \left(I_{d_2} - \eta \nabla^2_{yy} g(x, y^*(x))\right)\| < 1$ and $H(I_{d_2} - \eta \nabla^2_{yy} g(x, y^*(x)))$ is nonsingular. Therefore, we can obtain

$$\begin{aligned} A =& - \eta \left[I_{d_2} - H(I_{d_2} - \eta \nabla^2_{yy} g(x, y^*(x)))\right]^{-1} \\ &\cdot H \nabla^2_{yx} g(x, y^*(x)) \end{aligned} \quad (11)$$

Then, substituting A into the generalized gradient, we can obtain the following subset of the generalized gradient,

$$\begin{aligned} \bar{\partial} F(x) =& \{h | h = \nabla_x f(x, y^*(x)) - \eta \nabla^2_{xy} g(x, y^*(x)) H^\top \\ &\cdot \left[I_{d_2} - (I_{d_2} - \eta \nabla^2_{yy} g(x, y^*(x))) \cdot H^\top\right]^{-1} \\ &\cdot \nabla_y f(x, y^*(x)), H \in \partial \mathcal{P}_{\mathcal{Y}}(z^*)\}. \end{aligned} \quad (12)$$

Obviously, we have $\bar{\partial} F(x) \subset \partial F(x)$. Note that we can find that the hypergradient used in (Blondel et al., 2022; Bertrand et al., 2020; 2022) can be viewed as a specific element in our hypergradient. However, these methods need the projection operator to be differentiable, which is impractical in most cases. Therefore, we can conclude that our method can obtain the hypergradient in a more general case.

After obtaining the hypergradient, our next step is to design an algorithm to find the point x satisfying the condition

$$\min\{\|h\| : h \in \bar{\partial} F(x)\} \leq \epsilon, \quad (13)$$

such that x is a ϵ-Clarke stationary point (Clarke, 1990) which satisfies the condition

$$\min\{\|h\| : h \in \partial F(x)\} \leq \epsilon. \quad (14)$$

However, (Zhang et al., 2020) point out that finding an ϵ stationary points in nonsmooth nonconvex optimization can not be achieved by any finite-time algorithm given a fixed tolerance $\epsilon \in [0, 1)$. This suggests the need to rethink the definition of stationary points.

Inspired by (Zhang et al., 2020; Lin et al., 2022), we propose to consider the following δ-approximation generalized Jacobian.

Definition 3.4. Given a point $z \in \mathbb{R}^{d_2}$ and $\delta > 0$, the δ-approximation generalized Jacobian of a Lipschitz function of $\mathcal{P}_{\mathcal{Y}}(\cdot)$ at z is given by $\partial_\delta \mathcal{P}_{\mathcal{Y}}(z) := co\left(\bigcup_{z' \in \mathbb{B}_\delta(z)} \partial \mathcal{P}_{\mathcal{Y}}(z')\right)$.

The above approximation generalized Jacobian of $\mathcal{P}_{\mathcal{Y}}(\cdot)$ at z is the convex hull of all generalized Jacobians at points in a δ-ball around z, and it can be viewed as the extension of Goldstein subdifferential (Zhang et al., 2020; Lin et al., 2022; Goldstein, 1977). Then, substituting $\partial_\delta \mathcal{P}_{\mathcal{Y}}(z)$ into our hypergradient $\bar{\partial} F(x)$, we can obtain the following

approximation,

$$\begin{aligned}\bar{\partial}_\delta F(x) =&\{h|h = \nabla_x f(x, y^*(x)) - \eta\nabla^2_{xy} g(x, y^*(x))H^\top \\ &\cdot \left[I_{d_2} - (I_{d_2} - \eta\nabla^2_{yy} g(x, y^*(x))) \cdot H^\top\right]^{-1} \\ &\cdot \nabla_y f(x, y^*(x)), H \in \partial_\delta \mathcal{P}_\mathcal{Y}(z^*)\} \end{aligned} \tag{15}$$

Equipping with this hypergradient, even though we are unable to find an ϵ-stationary point, one could hope to find a point that is close to an ϵ-stationary point. This motivates us to adopt the following more refined notion

Definition 3.5. A point x is called (δ, ϵ)-stationary if $\min\left\{\|h\| : h \in \bar{\partial}_\delta F(x)\right\} \leq \epsilon$.

Note that if we can find a point x' at most distance δ away from x such that x' is ϵ-stationary, then we know x is (δ, ϵ)-stationary. However, the contrary is not true. We also have the following result

Lemma 3.6. *The set $\partial_\delta \mathcal{P}_\mathcal{Y}(z)$ converges as $\delta \downarrow 0$ as $\lim_{\delta\downarrow 0} \partial_\delta \mathcal{P}_\mathcal{Y}(z) = \partial \mathcal{P}_\mathcal{Y}(z)$.*

This result enables an intuitive framework for transforming the non-asymptotic analysis of convergence to (δ, ϵ)-stationary points to classical asymptotic results for finding ϵ-stationary points. Thus, we conclude that finding a (δ, ϵ) stationary point is a reasonable optimality condition for the lower-level constrained bilevel optimization problem.

3.2. Randomized Smoothing

Inspired by the strong ability of randomized smoothing to deal with nonsmooth problems, in this subsection, we use this method to handle the non-smoothness of the projection operator. Given a non-expansive projection operator (Moreau, 1965) $\mathcal{P}_\mathcal{Y}(z)$ and uniform distribution $\mathbb{P}$ on a unit ball in ℓ_2-norm, we define the smoothing function as $\mathcal{P}_{\mathcal{Y}\delta}(z) = \mathbb{E}_{u\sim\mathbb{P}}[\mathcal{P}_\mathcal{Y}(z+\delta u)]$. Then, we have the following proposition.

Proposition 3.7. *Let $\mathcal{P}_{\mathcal{Y}\delta}(z) = \mathbb{E}_{u\sim\mathbb{P}}[\mathcal{P}_\mathcal{Y}(z+\delta u)]$ where $\mathbb{P}$ is a uniform distribution on a unit ball in ℓ_2-norm. Since that $\mathcal{P}_\mathcal{Y}$ is non-expansive and each element of $\mathcal{P}_\mathcal{Y}$ is Lipschitz continuous with constant L_p, we have $\mathcal{P}_{\mathcal{Y}\delta}(z)$ is differentiable and 1-Lipschitz continuous with cd_2L_p/δ-Lipschitz gradient, where $c > 0$ is constant.*

Using this randomized smoothing function to replace the approximation generalized Jacobian in Eqn (15), we can approximate the hypergradient as follows,

$$\begin{aligned}&\nabla F_\delta(x) \\ =&\nabla_x f(x, y^*(x)) - \eta\nabla^2_{xy} g(x, y^*(x))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z^*)^\top \\ &\cdot \left[I_{d_2} - (I_{d_2} - \eta\nabla^2_{yy} g(x, y^*(x)))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z^*)^\top\right]^{-1} \\ &\cdot \nabla_y f(x, y^*(x)).\end{aligned}$$

For this hypergradient estimation, we have the following conclusion.

Lemma 3.8. *Under Assumptions 2.1, 2.2, we have $\nabla F_\delta(x)$ is Lipschitz continuous w.r.t x.*

Lemma 3.8 indicates that we can use the traditional analysis framework to discuss the convergence performance using $\nabla F_\delta(x)$. However, as we discussed above, our purpose is to find the (δ, ϵ)-stationary points of LCBO. Therefore, a new challenge arises, namely, to show the relation between $\nabla F_\delta(x)$ and $\bar{\partial}_\delta F(x)$. Here, we first give the relation between the δ-approximation generalized Jacobian and the Jacobian of the randomized smoothing function as follows.

Proposition 3.9. *We have $\nabla\mathcal{P}_{\mathcal{Y}\delta}(z) \in \partial_\delta \mathcal{P}_\mathcal{Y}(z)$ for any $z \in \mathbb{R}^{d_2}$.*

Using this result, we have $\nabla F_\delta(x) \in \bar{\partial}_\delta F(x)$. This resolves an important question and forms the basis for analyzing our method, which means that once we find a point satisfying the condition $\|\nabla F_\delta(x)\| \leq \epsilon$, then it is a (δ, ϵ)-stationary point.

3.3. Approximation of Hypergradient

In this subsection, we discuss how to approximate the hypergradient $\nabla F_\delta(x)$ in an efficient method.

Since obtaining the optimal solution $y^*(x)$ is usually time-consuming, one proper method is to replace $y^*(x)$ with y as an approximation as follows,

$$\begin{aligned}&\nabla f_\delta(x, y) = \nabla_x f(x, y) - \eta\nabla^2_{xy} g(x, y)\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)^\top \\ &\cdot \left[I_{d_2} - (I_{d_2} - \eta\nabla^2_{yy} g(x, y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)^\top\right]^{-1} \nabla_y f(x, y).\end{aligned}$$

where $z = y - \eta\nabla_y g(x, y)$. Then, we can use the Neumann series (Ghadimi & Wang, 2018; Meyer, 2000) to efficiently approximate the matrix inverse, since we have $\|(I_{d_2} - \eta\nabla^2_{yy} g(x, y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)^\top\| < 1$ if $\eta < 1/\mu_g$. Therefore, we can obtain the approximation of the hypergradient as follows,

$$\begin{aligned}&\bar{\nabla} f_\delta(x, y) = \nabla_x f(x, y) - \eta\nabla^2_{xy} g(x, y)\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)^\top \\ &\cdot \sum_{i=0}^{Q-1} \left((I_{d_2} - \eta\nabla^2_{yy} g(x, y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)^\top\right)^i \nabla_y f(x, y).\end{aligned}$$

In addition, since calculating $\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)$ is impractical, we can use the following unbiased estimator of the gradient $\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)$ as a replacement,

$$\bar{H}(z; u) = \sum_{i=1}^{d_2} \frac{1}{2\delta} \left(\mathcal{P}_\mathcal{Y}(z + \delta u_i) - \mathcal{P}_\mathcal{Y}(z - \delta u_i)\right) u_i^\top,$$

where $u_i \sim \mathbb{P}$. Note that, in the program, we can calculate $\bar{H}$ in parallel, thereby reducing the actual running time. For $\bar{H}(z; u)$, we have the following conclusion.

Lemma 3.10. *We have* $\mathbb{E}\left[\bar{H}(z;u)\right] = \nabla\mathcal{P}_{\mathcal{Y}\delta}(z)$ *and* $\mathbb{E}\left[\left\|\bar{H}(z;u) - \nabla\mathcal{P}_{\mathcal{Y}\delta}(z)\right\|^2\right] \leq 16\sqrt{2\pi}d_2L_p^2$.

To further reduce the complexity caused by calculating multiple Jacobian-vector products, we can introduce an additional stochastic layer on the finite sum. Specifically, assume we have a parameter $Q > 0$ and a collection of independent samples $\bar{\xi} := \{u^0, \cdots, u^{c(Q)}\}$, where $c(Q) \sim \mathcal{U}\{0, \cdots, Q-1\}$. Then, we can approximate the gradient as follows,

$$\bar{\nabla}f_\delta(x,y;\bar{\xi}) = \nabla_x f(x,y) - \eta Q\nabla^2_{xy}g(x,y)\bar{H}(z;u^0)^\top \cdot \prod_{i=1}^{c(Q)}\left((I_{d_2} - \eta\nabla^2_{yy}g(x,y))\bar{H}(z;u^i)^\top\right)\nabla_y f(x,y), \quad (16)$$

where we have used the the convention $\prod_{i=1}^0 A = I$. We can conclude that the bias of the gradient estimator $\bar{\nabla}f_\delta(x,y;\bar{\xi})$ decays exponentially fast with Q, as summarized below:

Lemma 3.11. *Under Assumptions 2.1, 2.2and Lemma 3.10, setting* $\frac{1}{\mu_g}(1 - \frac{1}{4(2\pi)^{1/4}\sqrt{d_2}L_p}) \leq \eta < \frac{1}{\mu_g}$, *for any* $x \in \mathbb{R}^{d_1}$, $y \in \mathcal{Y}$, *we have* $\left\|\nabla f_\delta(x,y) - \mathbb{E}[\bar{\nabla}f_\delta(x,y;\bar{\xi})]\right\| \leq \frac{C_{gxy}C_{fy}}{\mu_g}(1-\eta\mu_g)^Q$. *Furthermore, the variance of* $\bar{\nabla}f_\delta(x,y;\bar{\xi})$ *is bounded as* $\mathbb{E}\left[\left\|\bar{\nabla}f_\delta(x,y;\bar{\xi}) - \mathbb{E}\left[\bar{\nabla}f_\delta(x,y;\bar{\xi})\right]\right\|^2\right] \leq \sigma_f(d_2)$, *where* $\sigma_f(d_2)$ *is defined in Appendix I.*

Algorithm 1 DMLCBO

Input: Initialize $x_1 \in \mathcal{X}$, $y_1 \in \mathcal{Y}$, $v_1 = \nabla_y g(x_1,y_1)$, $w_1 = \bar{\nabla}f_\delta(x_1,y_1;\bar{\xi}_1)$, η_k, τ, γ, β, α, Q and η.

1: **for** $k = 1, \cdots, K$ **do**
2: Update $x_{k+1} = x_k - \frac{\eta_k\gamma}{\mathcal{P}_{[1/c_u,1/c_l]}(\sqrt{m_{2,k}}+G_0)}w_k$.
3: Update $y_{k+1} = (1-\eta_k)y_k + \eta_k\mathcal{P}_{\mathcal{Y}}(y_k - \frac{\tau}{\mathcal{P}_{[1/c_u,1/c_l]}(\sqrt{m_{1,k}}+G_0)}v_k)$
4: Calculate the hyper-gradient $\bar{\nabla}f_\delta(x_{k+1},y_{k+1};\bar{\xi}_{k+1})$ according to Eqn. (16).
5: Update $w_{k+1} = (1-\alpha)w_k + \alpha\bar{\nabla}f_\delta(x_{k+1},y_{k+1};\bar{\xi}_{k+1})$.
6: Update $v_{k+1} = (1-\beta)v_k + \beta\nabla_y g(x_{k+1},y_{k+1})$.
7: **end for**

Output: x_r where $r \in \{1, \cdots, K\}$ is uniformly sampled.

3.4. Double-momentum Method for Lower-level Constrained Bilevel Optimization

Equipped with the hypergradient $\bar{\nabla}f_\delta(x,y;\bar{\xi})$, our next endeavor is to design a *single-loop single-timescale* algorithm to solve the constrained bilevel optimization problem (2). Our main idea is to adopt the double-momentum-based method and adaptive step-size method developed in (Huang & Huang, 2021; Khanduri et al., 2021; Shi et al., 2022). Our algorithm is summarized in Algorithm 1. Since we use the double-momentum method to solve the lower-level constrained bilevel optimization problem, we denote our method as DMLCBO.

Define $\alpha \in (0,1)$ and $\beta \in (0,1)$. For the lower-level problem, we can utilize the following projected gradient method with the momentum-based gradient estimator and adaptive step size to update y,

$$\begin{aligned}\hat{y}_{k+1} =& \mathcal{P}_{\mathcal{Y}}(y_k - \frac{\tau}{\mathcal{P}_{[1/c_u,1/c_l]}(\sqrt{m_{1,k}}+G_0)}v_k),\\ y_{k+1} =& (1-\eta_k)y_k + \eta_k\hat{y}_{k+1},\\ v_{k+1} =& (1-\beta)v_k + \beta\nabla_y g(x_k,y_k),\end{aligned}$$

where $\eta_k > 0$, $\tau > 0$. $G_0 > 0$ is used to avoid to prevent the molecule from being equal to 0. Here, we initialize $v_1 = \nabla_y g(x_1,y_1)$. Similarly, for the upper-level problem, we can utilize the following gradient method with the momentum-based gradient estimator and adaptive step size to update x,

$$\begin{aligned}x_{k+1} =& x_k - \frac{\eta_k\gamma}{\mathcal{P}_{[1/c_u,1/c_l]}(\sqrt{m_{2,k}}+G_0)}w_k,\\ w_{k+1} =& (1-\alpha)w_k + \alpha\bar{\nabla}f_\delta(x_k,y_k;\bar{\xi}_k),\end{aligned}$$

and we initialize $w_1 = \bar{\nabla}f_\delta(x_1,y_1;\bar{\xi}_1)$. Note we set $m_{i,k+1} = 0.99\cdot m_{1,k} + 0.01\cdot\mathcal{G}^2$, where $\mathcal{G}$ denotes the gradient estimation, which is used in Adam (Kingma & Ba, 2014).

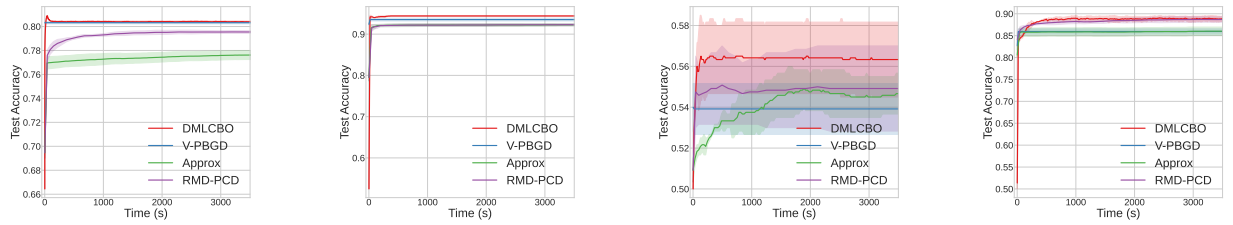
4. Convergence Analysis

In this section, we discuss the convergence performance of our DMLCBO (All the detailed proofs are presented in our Appendix). We follow the theoretical analysis framework in (Huang & Huang, 2021; Huang et al., 2020; Shi et al., 2022; Khanduri et al., 2021) (For easy understanding, we also provide a route map of the analysis in Appendix J).

Using all the assumptions and lemmas, we can obtain the following theorem (For ease of reading, some parameter Settings are omitted here. The specific parameters can be found in Appendix K.2):

Theorem 4.1. *Under Assumptions 2.1, 2.2 and Lemma 3.10, with* $\frac{1}{\mu_g}(1 - \frac{1}{4(2\pi)^{1/4}\sqrt{d_2}L_p}) \leq \eta < \frac{1}{\mu_g}$, $Q = \frac{1}{\mu_g\eta}\ln\frac{C_{gxy}C_{fy}K}{\mu_g}$, $0 \leq a \leq 2$, $\alpha = c_1\eta_k$, $\beta = c_2\eta_k$, $L_0 = \max(L_1(\frac{d_2}{\delta}), L_2(\frac{d_2}{\delta})) > 1$, $\Phi_1 = \mathbb{E}[F_\delta(x_1) + \frac{10L_0^a c_l}{\tau\mu_g c_u}\|y_1 - y^*(x_1)\|^2 + c_l(\|w_1 - \bar{\nabla}f_\delta(x_1,y_1) - R_1\|^2 + \|\nabla_y g(x_1,y_1) - v_1\|^2)]$, *and* $\eta_k = \frac{t}{(m+k)^{1/2}}$, $t > 0$, *we have*

$$\min\{\|h\| : h \in \bar{\partial}_\delta F(x_r)\} \leq \frac{4m^{1/4}\sqrt{G}}{\sqrt{Kt}} + \frac{4\sqrt{G}}{(Kt)^{1/4}}.$$



(a) CodRNA (b) MNIST 6vs9 (c) Madelon (d) FashionMNIST 1vs7

Figure 1: Test accuracy against training time of all the methods in data hyper-cleaning.

Table 2: Test accuracy (%) with standard variance of all the methods in data hyper-cleaning. (Higher is better.)

Datasets	DMLCBO	V-PBGD	Approx	RMD-PCD
CodRNA	**80.42** $\pm$ 0.05	80.32 $\pm$ 0.09	77.67 $\pm$ 0.38	79.43 $\pm$ 0.07
MNIST 6vs9	**94.24** $\pm$ 0.01	93.53 $\pm$ 0.34	92.29 $\pm$ 0.27	92.33 $\pm$ 0.29
Madelon	**56.33** $\pm$ 1.83	53.91 $\pm$ 1.25	54.75 $\pm$ 1.08	55.00 $\pm$ 2.17
FashionMNIST 1vs7	**88.85** $\pm$ 0.60	85.95 $\pm$ 0.60	86.03 $\pm$ 0.96	88.93 $\pm$ 0.68

where $G = \frac{\Phi_1 - \Phi^*}{\gamma c_l} + \frac{17t}{4K^2}(m+K)^{1/2} + \frac{4}{3tK^2}(m+K)^{3/2} + (m\sigma_f(d_2))t^2\ln(m+K)$*; the range of* c_1,c_2 γ, τ *and* m *are given in the appendix.*

Remark 4.2. Let $t = \mathcal{O}(1)$, $m = \mathcal{O}(1)$, $\sigma_f(d_2) = \mathcal{O}(d_2)$ and $\ln(m+K) = \tilde{\mathcal{O}}(1)$, we have $\sqrt{G} = \tilde{\mathcal{O}}((\frac{d_2}{\delta})^{a/2} + \sqrt{d_2})$. Thus, our proposed DMLCBO can converge to a (δ, ϵ) stationary point at the rate of $\tilde{\mathcal{O}}(((\frac{d_2}{\delta})^{a/2} + \sqrt{d_2})K^{-1/4})$ with properly choosing the hyper-parameters. Then, setting $a = 0$ and $\min\{\|h\| : h \in \bar{\partial}_\delta F(x_r)\} \leq \tilde{\mathcal{O}}(((\frac{d_2}{\delta})^{a/2} + \sqrt{d_2})K^{-1/4}) \leq \epsilon$, we have $K = \tilde{\mathcal{O}}(d_2^2\epsilon^{-4})$. Obviously, by setting $a = 0$, our method can avoid the influence of δ on the convergence iterations (Lin et al., 2022) and obtain the same convergence iteration number as the traditional bilevel optimization method without the Lipschitz assumption on the stochastic gradient estimation (Huang & Huang, 2021). Then, since we need to approximate $c(Q) \times d_2$ Jacobian-vector products and $\mathcal{O}(p)$ time to calculate the projection in each iteration, the average computational complexity of approximating hypergradient is $\mathcal{O}(Qd_2p + d_1)$. Therefore, the total computational complexity of our method is $\tilde{\mathcal{O}}((Qd_2p + d_1)d_2^2\epsilon^{-4})$. It is worth noting that since we approximate d_2 Jacobians in parallel, the overall execution time of our algorithm can be significantly reduced.

5. Experiments

In this section, we compare the performance of our method with SOTA methods for LCBO in two applications. (Detailed settings are given in our Appendix.)

5.1. Baselines

We compare our method with the following LCBO methods.

1. **V-PBGD**. The method proposed in (Shen & Chen, 2023) uses the value function method to solve the bilevel optimization problem.

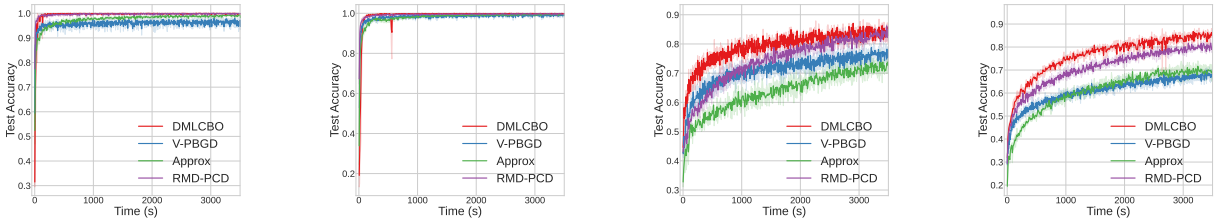
2. **RMD-PCD**. The method proposed in (Bertrand et al., 2022) which uses the reverse method to calculate the hyper-gradient.

3. **Approx**. The method proposed in (Pedregosa, 2016) solves a linear optimization problem to calculate the hypergradient.

We implement all the methods by Pytorch (Paszke et al., 2019). Since JaxOpt (Blondel et al., 2022) is implemented by JAX (Bradbury et al., 2018), for a fair comparison, we use Approx with the Jacobian calculating methods in (Martins & Astudillo, 2016; Djolonga & Krause, 2017; Blondel et al., 2020; Niculae & Blondel, 2017; Vaiter et al., 2013; Cherkaoui et al., 2020) as a replacement of JaxOpt, which uses the same method to calculate the hypergradient. We run all the methods 10 times on a PC with four 1080Ti GPUs.

5.2. Applications

Data hyper-cleaning. In this experiment, we evaluate the performance of all the methods in the application named data hyper-cleaning. In many real-world applications, the training set and testing set may have different distributions. To reduce the discrepancy between the two distributions, each data point will be given an additional importance weight, which is called data hyper-clean. This problem



(a) Omniglot 5 way 5 shot (b) Omniglot 10 way 5 shot (c) FC100 5 way 5 shot (d) FC100 10 way 5 shot

Figure 2: Test accuracy against training time of all the methods in meta-learning.

Table 3: Test accuracy (%) with standard variance in meta-learning. (Higher is better.)

Datasets	Problem	DMLCBO	V-PBGD	Approx	RMD-PCD
Omniglot	5 way 5 shot	**99.97** $\pm$ 0.04	95.67 $\pm$ 1.02	99.41 $\pm$ 0.25	99.69 $\pm$ 0.37
Omniglot	10 way 5 shot	**99.81** $\pm$ 0.19	98.43 $\pm$ 0.16	99.57 $\pm$ 0.17	99.29 $\pm$ 0.17
FC100	5 way 5 shot	**85.25** $\pm$ 0.01	78.14 $\pm$ 1.97	74.19 $\pm$ 1.35	84.75 $\pm$ 0.65
FC100	10 way 5 shot	**86.93** $\pm$ 0.89	67.18 $\pm$ 0.75	69.10 $\pm$ 0.85	81.18 $\pm$ 0.39

can be formulated as

$$\min_{x\in\mathbb{R}^{d_1}} \sum_{\mathcal{D}_{val}} \ell\left(y^*(x)^\top a_i, b_i\right)$$
$$s.t. \quad y^*(x) = \arg\min_{\|y\|_1\le r} \sum_{\mathcal{D}_{tr}} [\sigma(x)]_i \ell\left(y^\top a_i, b_i\right) + c\|y\|^2,$$

where $r > 0$, $\mathcal{D}_{tr}$ and $\mathcal{D}_{val}$ denote the training set and validation set respectively; (a_i, b_i) denotes the data point; $\sigma(\cdot) := 1/(1 + exp(-\cdot))$ is the Sigmoid function; $\ell(\cdot,\cdot)$ is the loss function; $c > 0$ is the regularization parameter used to ensure the lower-level problem to be strongly convex. In this experiment, an additional ℓ_1 is added to the lower-level problem to ensure the sparsity of the model.

Meta-learning. Meta-learning for few-shot learning is to learn a shared prior parameter across a distribution of tasks, such that a simple learning step with few-shot data based on the prior leads to a good adaptation to the task in the distribution. In particular, the training task $\mathcal{T}_i$ is sampled from distribution $P_{\mathcal{T}}$. Each task $\mathcal{T}_i$ is characterized by its training data $\mathcal{D}^i_{tr}$ and the test data $\mathcal{D}^i_{te}$. The upper level is to extract features from input data and multi-class SVM served as the base learner in the lower-level optimization to classify the data on its extracted features. This problem can be formulated as

$$\min_{\phi} \sum_{\mathcal{T}_i\sim P_{\mathcal{T}}} \mathcal{L}(y_i^*(x), \phi_x, \mathcal{D}^i_{te}) \tag{17}$$
$$s.t. \quad y_i^*(x) = \arg\min_{0\le y_i\le Ce} \frac{1}{2} y_i^\top G y_i - e^\top y_i$$

where ϕ_x denotes the network parameterized by x; e denotes the vector with all elements equal to 1; $G = K \odot (VV^\top)$, $K_{ij} = \phi(a_i)^\top \phi(a_j)$, $V_{ij} = \frac{1}{k-1}$ if $b_i \neq j$ otherwise $V_{ij} = 1$, $(a_j, b_j) \in \mathcal{D}^i_{tr}$, and K denotes the number of classes.

5.3. Results

We have presented the test accuracy results for all methods in Tables 2 and 3, and visualized the testing performance as a function of training time in Figures 1 and 2. Upon closer examination of the data presented in Tables 2 and 3, we can find that our method achieves similar performances in some cases and sometimes gets results better than other methods. One possible reason is that RMD-PCD, Approx, and V-PBGD highly depend on the solutions to the lower-level problem. Once the approximated solution is not good enough then they may obtain bad performance. From all these results, we can conclude that our DMLCBO outperforms other methods in terms of effectiveness.

6. Conclusion

In this paper, without using the restive assumptions on $y^*(x)$, we propose a new method to derive the hypergradient for LCBO. Then, we leverage randomized smoothing to approximate the hypergradient. Then, using our new hypergradient approximation, we propose a *single-loop single-timescale* algorithm based on the double-momentum method and adaptive step size method, which updates the lower- and upper-level variables simultaneously. Theoretically, we prove our methods can converge to the (δ, ϵ)-stationary point with $\tilde{\mathcal{O}}(d_2^2\epsilon^{-4})$. The experimental results in data hypercleaning and meta-learning demonstrate the effectiveness of our proposed method.

Acknowledgements

Bin Gu was supported by the Natural Science Foundation of China under Grant No.62076138. Yi Chang was supported by the Natural Science Foundation of China under Grant No.U2341229 and the National Key R&D Program of China under Grant No.2023YFF0905400.

Impact Statement

This paper presents work on the theoretical analysis of the bilevel optimization problem. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Bard, J. F. *Practical bilevel optimization: algorithms and applications*. Springer Science & Business Media, 2013.

Bergstra, J. S., Bardenet, R., Bengio, Y., and Kégl, B. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, pp. 2546–2554, 2011.

Bertrand, Q., Klopfenstein, Q., Blondel, M., Vaiter, S., Gramfort, A., and Salmon, J. Implicit differentiation of lasso-type models for hyperparameter optimization. In *International Conference on Machine Learning*, pp. 810–821. PMLR, 2020.

Bertrand, Q., Klopfenstein, Q., Massias, M., Blondel, M., Vaiter, S., Gramfort, A., Salmon, J., Chevalier, J., Nguyen, T., Thirion, B., et al. Implicit differentiation for fast hyperparameter selection in non-smooth convex learning. *Journal of Machine Learning Research*, 23(149):1–43, 2022.

Bertsekas, D. P. On penalty and multiplier methods for constrained minimization. *SIAM Journal on Control and Optimization*, 14(2):216–235, 1976.

Blondel, M., Teboul, O., Berthet, Q., and Djolonga, J. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, pp. 950–959. PMLR, 2020.

Blondel, M., Berthet, Q., Cuturi, M., Frostig, R., Hoyer, S., Llinares-López, F., Pedregosa, F., and Vert, J.-P. Efficient and modular implicit differentiation. *Advances in Neural Information Processing Systems*, 35:5230–5242, 2022.

Bolte, J., Le, T., Pauwels, E., and Silveti-Falls, T. Nonsmooth implicit differentiation for machine-learning and optimization. *Advances in neural information processing systems*, 34:13537–13549, 2021.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Chen, T., Sun, Y., and Yin, W. A single-timescale stochastic bilevel optimization method. *arXiv preprint arXiv:2102.04671*, 2021.

Cherkaoui, H., Sulam, J., and Moreau, T. Learning to solve tv regularised problems with unrolled algorithms. *Advances in Neural Information Processing Systems*, 33: 11513–11524, 2020.

Clarke, F. H. *Optimization and nonsmooth analysis*. SIAM, 1990.

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

Djolonga, J. and Krause, A. Differentiable learning of submodular models. *Advances in Neural Information Processing Systems*, 30, 2017.

Feurer, M., Springenberg, J., and Hutter, F. Initializing bayesian hyperparameter optimization via meta-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, pp. 1128–1135, 2015.

Franceschi, L., Donini, M., Frasconi, P., and Pontil, M. Forward and reverse gradient-based hyperparameter optimization. In *International Conference on Machine Learning*, pp. 1165–1173. PMLR, 2017.

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.

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

Goldstein, A. Optimization of lipschitz continuous functions. *Mathematical Programming*, 13:14–22, 1977.

Guo, Z., Hu, Q., Zhang, L., and Yang, T. Randomized stochastic variance-reduced methods for multi-task stochastic bilevel optimization. *arXiv preprint arXiv:2105.02266*, 2021.

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. and Huang, H. Biadam: Fast adaptive bilevel optimization methods. *arXiv preprint arXiv:2106.11396*, 2021.

Huang, F., Gao, S., Pei, J., and Huang, H. Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization. *arXiv preprint arXiv:2008.08170*, 2020.

Ji, K., Yang, J., and Liang, Y. Bilevel optimization: Nonasymptotic analysis and faster algorithms. In *International Conference on Machine Learning (ICML)*, 2021.

Khanduri, P., Zeng, S., Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in neural information processing systems*, 34:30271–30283, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Konda, V. R. and Tsitsiklis, J. N. Actor-critic algorithms. In *Advances in neural information processing systems*, pp. 1008–1014. Citeseer, 2000.

Lin, T., Zheng, Z., and Jordan, M. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. *Advances in Neural Information Processing Systems*, 35:26160–26175, 2022.

Liu, R., Liu, Y., Zeng, S., and Zhang, J. Towards gradient-based bilevel optimization with non-convex followers and beyond. *Advances in Neural Information Processing Systems*, 34:8662–8675, 2021.

Martins, A. and Astudillo, R. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pp. 1614–1623. PMLR, 2016.

Meyer, C. D. *Matrix analysis and applied linear algebra*, volume 71. Siam, 2000.

Moreau, J.-J. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

Niculae, V. and Blondel, M. A regularized framework for sparse and structured neural attention. *Advances in neural information processing systems*, 30, 2017.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Pedregosa, F. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pp. 737–746. PMLR, 2016.

Rajeswaran, A., Finn, C., Kakade, S., and Levine, S. Meta-learning with implicit gradients. *arXiv preprint arXiv:1909.04630*, 2019.

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

Shi, W. and Gu, B. Improved penalty method via doubly stochastic gradients for bilevel hyperparameter optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):9621–9629, May 2021. doi: 10.1609/aaai.v35i11.17158. URL https://ojs.aaai.org/index.php/AAAI/article/view/17158.

Shi, W., Gao, H., and Gu, B. Gradient-free method for heavily constrained nonconvex optimization. In *International Conference on Machine Learning*, pp. 19935–19955. PMLR, 2022.

Tsaknakis, I., Khanduri, P., and Hong, M. An implicit gradient-type method for linearly constrained bilevel problems. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5438–5442, 2022. doi: 10.1109/ICASSP43922.2022.9747013.

Vaiter, S., Deledalle, C.-A., Peyré, G., Dossal, C., and Fadili, J. Local behavior of sparse analysis regularization: Applications to risk estimation. *Applied and Computational Harmonic Analysis*, 35(3):433–451, 2013.

Van Leeuwen, T. and Aravkin, A. Y. Variable projection for nonsmooth problems. *SIAM journal on scientific computing*, 43(5):S249–S268, 2021.

Xiao, Q., Shen, H., Yin, W., and Chen, T. Alternating projected sgd for equality-constrained bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 987–1023. PMLR, 2023.

Xu, S. and Zhu, M. Efficient gradient approximation method for constrained bilevel optimization. *arXiv preprint arXiv:2302.01970*, 2023.

Zhang, J., Lin, H., Jegelka, S., Sra, S., and Jadbabaie, A. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, pp. 11173–11182. PMLR, 2020.

Appendix

A. Experimental Setting

Data hyper-cleaning. In this experiment, we evaluate all the methods on the datasets MNIST, FashionMNIST, CodRNA, and Madelon [1]. For MNIST and FashionMNIST, we choose two classes to conduct a binary classification. In addition, we flip 30% of the labels in the training set as the noisy data. We set $r = 1$ for all the datasets. For our method, we search the step size from the set $\{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. For other methods, we search the step size from the set $\{10, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. Following the default setting in (Ji et al., 2021), we set $Q = 3$ and $\eta = 0.5$ for our method. In addition, we set $\eta_k = 1/\sqrt{100 + k}$, $c_1 = 10$ and $c_2 = 10$ for our method. For V-PBGD, RMD-PCD, and Approx, following the setting in (Pedregosa, 2016), we set the inner iteration number at 100. We run all the methods for 10000 iterations and evaluate the test accuracy for every 100 iteration. We set $\delta = 1e - 6$.

Meta learning. In this experiment, we evaluate all the methods on the datasets Omniglot and FC100. We set $C = 10$. We use a network with four convolution-ELU-Maxpooling layers to extract the features that output a $\mathbb{R}^{100}$ vector as the feature. For our method, we search the step size from the set $\{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. For other methods, we search the step size from the set $\{10, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. Following the default setting in (Ji et al., 2021), we set $Q = 3$ and $\eta = 0.5$ for our method. In addition, we set $\eta_k = 1/\sqrt{100 + k}$, $c_1 = 10$ and $c_2 = 10$ for our method. For V-PBGD, RMD-PCD, and Approx, we also set the inner iteration number at 3. We set $\delta = 1e - 6$.

B. Ablation Study

In this section, we conduct ablation experiments on the hyper-parameters Q and η. To control the variables, we explore the effect of each hyper-parameter while keeping the other hyper-parameters as default as shown in the experimental setups. We search the step size of both x and y. We present the results in Tables 4, 5. We can find that increasing Q will lead to a long training time. In addition, setting $Q = 1$ usually leads to the worst results, which is because setting $Q = 1$ means ignoring the inverse of the Hessian matrix in our hypergradient and may not converge to our stationary point.

Table 4: Test accuracy (%) of our method with different Q in data hyper-cleaning. (Higher is better.)

Datasets	Q=1	Q=3	Q=5	Q=7
CodRNA	80.25 ± 0.07	80.42 ± 0.05	80.55 ± 0.03	**80.56** ± 0.04
MNIST 6vs9	93.65 ± 0.01	94.24 ± 0.01	94.15 ± 0.01	**94.35** ± 0.01
Madelon	55.23 ± 1.22	56.33 ± 1.83	**56.67** ± 1.24	56.23 ± 1.56
FashionMNIST 1vs7	88.25 ± 0.56	**88.85** ± 0.60	88.75 ± 0.40	88.55 ± 0.33

Table 5: Test accuracy (%) of our method with different η in data hyper-cleaning. (Higher is better.)

Datasets	$\eta = 0.1$	$\eta = 0.5$	$\eta = 1$
CodRNA	80.45 ± 0.04	80.42 ± 0.05	80.33 ± 0.06
MNIST 6vs9	94.65 ± 0.01	94.24 ± 0.01	94.35 ± 0.01
Madelon	56.23 ± 1.14	56.33 ± 1.83	56.23 ± 1.56
FashionMNIST 1vs7	88.65 ± 0.67	88.85 ± 0.60	88.67 ± 0.23

C. Proof of Lemma 3.3

Proof. Here, we follow the proof in (Van Leeuwen & Aravkin, 2021). For given x_1 and x_2, we have the corresponding unique optimal solutions $y^*(x_1)$ and $y^*(x_2)$. For the constrained lower-level problem, we have the following optimal conditions

$$0 \in \nabla_y g(x_1, y^*(x_1)) + \partial\delta_{\mathcal{Y}}(y^*(x_1)), \quad 0 \in \nabla_y g(x_2, y^*(x_2)) + \partial\delta_{\mathcal{Y}}(y^*(x_2)) \tag{18}$$

where $\delta_{\mathcal{Y}}(\cdot)$ is the indicator function of the constriant.

[1] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

Since g is strongly convex for any given $\tilde{x}$ and $\mathcal{Y}$ is convex, we have

$$\mu_g\|y^*(x_1)-y^*(x_2)\|^2 \leq \langle(\nabla_y g(\tilde{x},y^*(x_1))+h_1)-(\nabla_y g(\tilde{x},y^*(x_2))+h_2)\,,y^*(x_1)-y^*(x_2)\rangle \tag{19}$$

where $h_1 \in \partial\delta_{\mathcal{Y}}(y^*(x_1))$ and $h_2 \in \partial\delta_{\mathcal{Y}}(y^*(x_2))$. Make the particular choices $h_1 = -\nabla_y g(x_1,y^*(x_1)) \in \partial\delta_{\mathcal{Y}}(y^*(x_1))$ and $h_2 = -\nabla_y g(x_2,y^*(x_2)) \in \partial\delta_{\mathcal{Y}}(y^*(x_2))$, we have

$$\begin{aligned}
&\mu_g\|y^*(x_1)-y^*(x_2)\|^2\\
\leq&\langle\nabla_y g(\tilde{x},y^*(x_1))-\nabla_y g(x_1,y^*(x_1)),y^*(x_1)-y^*(x_2)\rangle\\
&+\langle\nabla_y g(x_2,y^*(x_2))-\nabla_y g(\tilde{x},y^*(x_2)),y^*(x_1)-y^*(x_2)\rangle
\end{aligned} \tag{20}$$

Then, setting $\tilde{x}=x_1$, we have

$$\begin{aligned}
&\mu_g\|y^*(x_1)-y^*(x_2)\|^2\\
\leq&\langle\nabla_y g(x_2,y^*(x_2))-\nabla_y g(x_1,y^*(x_2)),y^*(x_1)-y^*(x_2)\rangle\\
\leq&\|\nabla_y g(x_2,y^*(x_2))-\nabla_y g(x_1,y^*(x_2))\|\|y^*(x_1)-y^*(x_2)\|\\
\leq&L_g\|x_1-x_2\|\|y^*(x_1)-y^*(x_2)\|
\end{aligned} \tag{21}$$

where the last inequality is due to Assumption 2.2. Rearrange above inequality, we have

$$\|y^*(x_1)-y^*(x_2)\| \leq \frac{L_g}{\mu_g}\|x_1-x_2\| \tag{22}$$

That completes the proof. □

D. Proof of Lemma 3.6

Proof. First, we show that the limit exists. By Lipschitzness of the projection operator and Jenson inequality, we know that $\partial_\delta\mathcal{P}_{\mathcal{Y}}(z)$ lies in a bounded ball with radius 1. For any sequence of δ_k with $\delta_k \downarrow 0$, we know that $\partial_{\delta_{k+1}}\mathcal{P}_{\mathcal{Y}}(z) \subset \partial_{\delta_k}\mathcal{P}_{\mathcal{Y}}(z)$. Therefore, the limit exists by the monotone convergence theorem.

Next, we show that $\lim_{\delta\downarrow 0}\partial_\delta\mathcal{P}_{\mathcal{Y}}(z) = \partial\mathcal{P}_{\mathcal{Y}}(z)$. According to the Proposition 2.6.2 in (Clarke, 1990), we have

$$\partial\mathcal{P}_{\mathcal{Y}}(z) = \bigcap_{\delta>0}\bigcup_{z'\in\mathbb{B}_\delta(z)}\partial\mathcal{P}_{\mathcal{Y}}(z'). \tag{23}$$

Then, using the fact

$$\bigcup_{z'\in\mathbb{B}_\delta(z)}\partial\mathcal{P}_{\mathcal{Y}}(z') \subseteq co\left(\bigcup_{z'\in\mathbb{B}_\delta(z)}\partial\mathcal{P}_{\mathcal{Y}}(z')\right) = \partial_\delta\mathcal{P}_{\mathcal{Y}}(z), \tag{24}$$

we have $\partial\mathcal{P}_{\mathcal{Y}}(z) \subseteq \lim_{\delta\downarrow 0}\partial_\delta\mathcal{P}_{\mathcal{Y}}(z)$.

Using the upper semicontinuous of $\partial\mathcal{P}_{\mathcal{Y}}(z)$ at z (Clarke, 1990), we have that for any $\epsilon>0$, there exists $\delta>0$ such that

$$\bigcup_{z'\in\mathbb{B}_\delta(z)}\partial\mathcal{P}_{\mathcal{Y}}(z') \subseteq \partial\mathcal{P}_{\mathcal{Y}}(z)+\epsilon B \tag{25}$$

Then, by convexity of $\partial\mathcal{P}_{\mathcal{Y}}(z)$ and ϵB, we know that their Minknowski sum $\partial\mathcal{P}_{\mathcal{Y}}(z)+\epsilon B$ isconvex. Therefore, we conclude that for any $\epsilon>0$, there exists $\delta>0$ such that

$$\partial_\delta\mathcal{P}_{\mathcal{Y}}(z) = co\left(\bigcup_{z'\in\mathbb{B}_\delta(z)}\partial\mathcal{P}_{\mathcal{Y}}(z')\right) \subseteq \partial\mathcal{P}_{\mathcal{Y}}(z)+\epsilon B \tag{26}$$

Therefore, we have $\lim_{\delta\downarrow 0}\partial_\delta\mathcal{P}_{\mathcal{Y}}(z) \subseteq \partial\mathcal{P}_{\mathcal{Y}}(z)$. Finally, we obtain $\lim_{\delta\downarrow 0}\partial_\delta\mathcal{P}_{\mathcal{Y}}(z) = \partial\mathcal{P}_{\mathcal{Y}}(z)$. □

E. Proof of Proposition 3.7

Proof. Let $u \in \mathbb{R}^d$ denote a random variable distributed uniformly on $\mathbb{B}_1(0)$. Since $\mathcal{P}_{\mathcal{Y}}(\cdot)$ is 1-Lipschitz, we have

$$\|\mathcal{P}_{\mathcal{Y}\delta}(z) - \mathcal{P}_{\mathcal{Y}}(z)\| = \|\mathbb{E}[\mathcal{P}_{\mathcal{Y}}(z+\delta u) - \mathcal{P}_{\mathcal{Y}}(z)]\| \leq \delta \cdot 1 \cdot \mathbb{E}[\|u\|] = \delta \tag{27}$$

Since $\mathcal{P}_{\mathcal{Y}}(z)$ is 1-Lipschitz, we have

$$\|\mathcal{P}_{\mathcal{Y}\delta}(z) - \mathcal{P}_{\mathcal{Y}\delta}(z')\| = \|\mathbb{E}[\mathcal{P}_{\mathcal{Y}}(z+\delta u) - \mathcal{P}_{\mathcal{Y}}(z'+\delta u)]\| \leq \|\mathbb{E}[\|z-z'\|]\| = \|z-z'\| \tag{28}$$

For the Jacobian matrix, we have

$$\nabla \mathcal{P}_{\mathcal{Y}\delta}(z) = \begin{bmatrix} (\nabla \mathcal{P}^1_{\mathcal{Y}\delta}(z))^\top \\ \vdots \\ (\nabla \mathcal{P}^{d_2}_{\mathcal{Y}\delta}(z))^\top \end{bmatrix} \tag{29}$$

For each element $\nabla \mathcal{P}^i_{\mathcal{Y}\delta}(z)$, $(i \in [1, \cdots, d_2])$, according to Theorem 3.1 in (Lin et al., 2022), we have $\nabla \mathcal{P}^i_{\mathcal{Y}\delta}(z) = \mathbb{E}[\nabla \mathcal{P}^i_{\mathcal{Y}}(z+\delta u)]$. Therefore, we can easily obtain $\nabla \mathcal{P}_{\mathcal{Y}\delta}(z) = \mathbb{E}[\nabla \mathcal{P}_{\mathcal{Y}}(z+\delta u)]$. Then, according to (Lin et al., 2022), we have

$$\begin{aligned} \|\nabla \mathcal{P}_{\mathcal{Y}\delta}(z) - \nabla \mathcal{P}_{\mathcal{Y}\delta}(z')\| &\leq \|\nabla \mathcal{P}_{\mathcal{Y}\delta}(z) - \nabla \mathcal{P}_{\mathcal{Y}\delta}(z')\|_F = \left(\sum_{i=1}^{d_2} \|\nabla \mathcal{P}^i_{\mathcal{Y}\delta}(z) - \nabla \mathcal{P}^i_{\mathcal{Y}\delta}(z')\|^2 \right)^{1/2} \\ &\leq \left(\sum_{i=1}^{d_2} (\frac{cL_p\sqrt{d_2}}{\delta}\|z-z'\|)^2 \right)^{1/2} = \frac{cd_2L_p}{\delta}\|z-z'\| \end{aligned} \tag{30}$$

where $c > 0$ is a constant. □

F. Proof of Proposition 3.9

Proof. According to Theorem 3.1 in (Lin et al., 2022), for each element $\nabla \mathcal{P}^i_{\mathcal{Y}\delta}(z)$ in $\nabla \mathcal{P}_{\mathcal{Y}\delta}(z)$, we have $\nabla \mathcal{P}^i_{\mathcal{Y}\delta}(z) \in \partial_\delta \mathcal{P}^i_{\mathcal{Y}}(z)$. Therefore, we have $\nabla \mathcal{P}_{\mathcal{Y}\delta}(z) \in \partial_\delta \mathcal{P}_{\mathcal{Y}}(z)$ □

G. Lipschitz Continuousness of $\nabla F_\delta(x)$

Here, we prove $\nabla F_\delta(x)$ is Lipschitz continuous. We first give several useful lemmas.

Lemma G.1. ***(Lipschitz continuous of the approximation of hypergradient on x and y.)*** *Under Assumptions 2.1, 2.2, $\nabla f_\delta(x,y)$ is Lipschitz continuous on $y \in \mathcal{Y}$ and $x \in \mathcal{X}$, respectively, such that we have*

$$\|\nabla f_\delta(x,y_1) - \nabla f_\delta(x,y_2)\| \leq L_1(\frac{d_2}{\delta})\|y_1 - y_2\| \tag{31}$$

$$\|\nabla f_\delta(x_1,y) - \nabla f_\delta(x_2,y)\| \leq L_2(\frac{d_2}{\delta})\|x_1 - x_2\| \tag{32}$$

where $L_1(\frac{d_2}{\delta}) = L_f + \frac{L_{gxy}C_{fy}}{\mu_g} + \frac{C_{gxy}C_{fy}}{\mu_g}(\frac{cd_2}{\delta} + \eta\frac{cd_2}{\delta}L_g) + C_{gxy}C_{fy}\frac{1}{\eta\mu_g^2}(\frac{cd_2}{\delta} + \eta\frac{cd_2}{\delta}L_g)(1-\eta\mu_g) + \frac{L_{gyy}}{\mu_g^2} + \frac{C_{gxy}}{\mu_g}L_f$ *and* $L_2(\frac{d_2}{\delta}) = L_f + \frac{L_{gxy}C_{fy}}{\mu_g} + \frac{C_{gxy}C_{fy}}{\mu_g}\eta\frac{cd_2}{\delta}L_g + C_{gxy}C_{fy}\left(\eta\frac{1}{\mu_g^2}\frac{cd_2}{\delta}L_g(1-\eta\mu_g) + \frac{L_{gyy}}{\mu_g^2}\right) + \frac{C_{gxy}}{\mu_g}L_f$

Proof. Using the definition of $\nabla f_\delta(x,y)$, $z_1 = y_1 - \nabla_y g(x,y_1)$ and $z_2 = y_2 - \nabla_y g(x,y_2)$, we have

$$\begin{aligned}
&\|\nabla f_\delta(x,y_1) - \nabla f_\delta(x,y_2)\| \\
=&\|\nabla_x f(x,y_1) - \eta\nabla^2_{xy} g(x,y_1)\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)^\top \left[(I_{d_2} - (I_{d_2} - \eta\nabla^2_{yy} g(x,y_1))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)^\top\right]^{-1}\nabla_y f(x,y_1) \\
&- \nabla_x f(x,y_2) + \eta\nabla^2_{xy} g(x,y_2)\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)^\top \left[(I_{d_2} - (I_{d_2} - \eta\nabla^2_{yy} g(x,y_2))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)^\top\right]^{-1}\nabla_y f(x,y_2)\| \\
\le&\|\nabla_x f(x,y_1) - \nabla f(x,y_2)\| \\
&+ \eta\|\nabla^2_{xy} g(x,y_2) - \nabla^2_{xy} g(x,y_1)\|\|\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)\|\|\left[I_{d_2} - (I_{d_2} - \eta\nabla^2_{yy} g(x,y_2))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)^\top\right]^{-1}\|\|\nabla_y f(x,y_2)\| \\
&+ \eta\|\nabla^2_{xy} g(x,y_1)\|\|\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2) - \nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)\|\|\left[I_{d_2} - (I_{d_2} - \eta\nabla^2_{yy} g(x,y_2))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)^\top\right]^{-1}\|\|\nabla_y f(x,y_2)\| \\
&+ \eta\|\nabla^2_{xy} g(x,y_1)\|\|\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)\|\|\left[I_{d_2} - (I_{d_2} - \eta\nabla^2_{yy} g(x,y_2))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)^\top\right]^{-1} \\
&- \left[I_{d_2} - (I_{d_2} - \eta\nabla^2_{yy} g(x,y_1))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)^\top\right]^{-1}\|\|\nabla_y f(x,y_2)\| \\
&+ \eta\|\nabla^2_{xy} g(x,y_1)\|\|\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)\|\|\left[I_{d_2} - (I_{d_2} - \eta\nabla^2_{yy} g(x,y_1))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)^\top\right]^{-1}\|\|\nabla_y f(x,y_2) - \nabla_y f(x,y_1)\|
\end{aligned} \tag{33}$$

We have

$$\|\nabla_x f(x,y_1) - \nabla_x f(x,y_2)\| \le L_f\|y_1 - y_2\|, \tag{34}$$

$$\|\nabla_y f(x,y_1) - \nabla_y f(x,y_2)\| \le L_f\|y_1 - y_2\|, \tag{35}$$

$$\|\nabla^2_{xy} g(x,y_2) - \nabla^2_{xy} g(x,y_1)\| \le L_{gxy}\|y_2 - y_1\|, \tag{36}$$

$$\|\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)\| \le 1, \quad \|\nabla f(x,y_2)\| \le C_{fy}, \quad \|\nabla^2_{xy} g(x,y_1)\| \le C_{gxy} \tag{37}$$

Since $\|(I_{d_2} - \eta\nabla^2_{yy} g(x,y_2))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)^\top\| \le \|\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)\|\|I_{d_2} - \eta\nabla^2_{yy} g(x,y_2)\| \le 1 - \eta\mu_g \le 1$ we have

$$\begin{aligned}
&\|\left[I_{d_2} - (I_{d_2} - \eta\nabla^2_{yy} g(x,y_2))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)^\top\right]^{-1}\| \\
\le&\frac{1}{1 - \|(I_{d_2} - \eta\nabla^2_{yy} g(x,y_2))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)^\top\|} \\
\le&\frac{1}{\eta\mu_g}
\end{aligned} \tag{38}$$

$$\begin{aligned}
&\|\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2) - \nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)\| \\
\le&\frac{cd_2}{\delta}\|z_2 - z_1\| \\
=&\frac{cd_2}{\delta}\|y_2 - \eta\nabla_y g(x,y_2) - y_1 + \eta\nabla_y g(x,y_1)\| \\
\le&\frac{cd_2}{\delta}\|y_2 - y_1\| + \eta\frac{cd_2}{\delta}\|\nabla_y g(x,y_1) - \eta\nabla_y g(x,y_2)\| \\
\le&(\frac{cd_2}{\delta} + \eta\frac{cd_2}{\delta}L_g)\|y_2 - y_1\|
\end{aligned} \tag{39}$$

Using the inequality $\|H_2^{-1} - H_1^{-1}\| \le \|H_1^{-1}(H_1 - H_2)H_2^{-1}\| \le \|H_1^{-1}\|\|H_1 - H_2\|\|H_2^{-1}\|$, we have

$$\begin{aligned}
&\|\left[I_{d_2} - (I_{d_2} - \eta\nabla^2_{yy} g(x,y_2))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)^\top\right]^{-1} - \left[I_{d_2} - (I_{d_2} - \eta\nabla^2_{yy} g(x,y_1))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)^\top\right]^{-1}\| \\
\le&\frac{1}{\eta^2\mu_g^2}\|(I_{d_2} - \eta\nabla^2_{yy} g(x,y_2))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)^\top - (I_{d_2} - \eta\nabla^2_{yy} g(x,y_1))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)^\top\| \\
\le&\frac{1}{\eta^2\mu_g^2}\|\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2) - \nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)\|\|I_{d_2} - \eta\nabla^2_{yy} g(x,y_2)\| \\
&+ \frac{1}{\eta^2\mu_g^2}\|\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)\|\|I_{d_2} - \eta\nabla^2_{yy} g(x,y_2) - I_{d_2} + \eta\nabla^2_{yy} g(x,y_1)\| \\
\le&\left(\frac{1}{\eta^2\mu_g^2}(\frac{cd_2}{\delta} + \eta\frac{cd_2}{\delta}L_g)(1 - \eta\mu_g) + \frac{L_{gyy}}{\eta\mu_g^2}\right)\|y_2 - y_1\|
\end{aligned} \tag{40}$$

Therefore, using the above inequalities, we obtain

$$\begin{aligned}
&\|\nabla f_\delta(x,y_1)-\nabla f_\delta(x,y_2\nabla f_\delta(x,y))\| \\
\leq& L_f+\frac{L_{gxy}C_{fy}}{\mu_g}+\frac{C_{gxy}C_{fy}}{\mu_g}(\frac{cd_2}{\delta}+\eta\frac{cd_2}{\delta}L_g) \\
&+C_{gxy}C_{fy}\left(\frac{1}{\eta\mu_g^2}(\frac{cd_2}{\delta}+\eta\frac{cd_2}{\delta}L_g)(1-\eta\mu_g)+\frac{L_{gyy}}{\mu_g^2}\right)+\frac{C_{gxy}}{\mu_g}L_f\|y_2-y_1\|
\end{aligned} \tag{41}$$

For the second statement, let $z_1=y-\nabla_y g(x_1,y)$ and $z_2=y-\nabla_y g(x_2,y)$, we have

$$\begin{aligned}
&\|\nabla f_\delta(x_1,y)-\nabla f_\delta(x_2,y)\| \\
=&\|\nabla_x f(x_1,y)-\eta\nabla^2_{xy}g(x_1,y)\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)^\top\left[I_{d_2}-(I_{d_2}-\eta\nabla^2_{yy}g(x_1,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)^\top\right]^{-1}\nabla_y f(x_1,y) \\
&-\nabla_x f(x_2,y)+\eta\nabla^2_{xy}g(x_2,y)\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)^\top\left[I_{d_2}-(I_{d_2}-\eta\nabla^2_{yy}g(x_2,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)^\top\right]^{-1}\nabla_y f(x_2,y)\| \\
\leq&\|\nabla_x f(x_1,y)-\nabla f(x_2,y)\| \\
&+\eta\|\nabla^2_{xy}g(x_2,y)-\nabla^2_{xy}g(x_1,y)\|\|\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)\|\|\left[I_{d_2}-(I_{d_2}-\eta\nabla^2_{yy}g(x_2,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)^\top\right]^{-1}\|\|\nabla_y f(x_2,y)\| \\
&+\eta\|\nabla^2_{xy}g(x_1,y)\|\|\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)-\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)\|\|\left[I_{d_2}-(I_{d_2}-\eta\nabla^2_{yy}g(x_2,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)^\top\right]^{-1}\|\|\nabla_y f(x_2,y)\| \\
&+\eta\|\nabla^2_{xy}g(x_1,y)\|\|\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)\|\|\left[I_{d_2}-(I_{d_2}-\eta\nabla^2_{yy}g(x_2,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)^\top\right]^{-1} \\
&-\left[I_{d_2}-(I_{d_2}-\eta\nabla^2_{yy}g(x_1,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)^\top\right]^{-1}\|\|\nabla_y f(x_2,y)\| \\
&+\eta\|\nabla^2_{xy}g(x_1,y)\|\|\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)\|\|\left[I_{d_2}-(I_{d_2}-\eta\nabla^2_{yy}g(x_1,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)^\top\right]^{-1}\|\|\nabla_y f(x_2,y)-\nabla_y f(x_1,y)\|
\end{aligned} \tag{42}$$

We have

$$\|\nabla_x f(x_1,y)-\nabla_x f(x_2,y)\|\leq L_f\|x_2-x_1\| \tag{43}$$
$$\|\nabla_y f(x_1,y)-\nabla_y f(x_2,y)\|\leq L_f\|x_2-x_1\| \tag{44}$$
$$\|\nabla^2_{xy}g(x_2,y)-\nabla^2_{xy}g(x_1,y)\|\leq L_{gxy}\|x_2-x_1\| \tag{45}$$

$$\begin{aligned}
&\|\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)-\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)\| \\
\leq&\frac{cd_2}{\delta}\|z_2-z_1\| \\
=&\frac{cd_2}{\delta}\|y-\eta\nabla_y g(x_2,y)-y+\eta\nabla_y g(x_1,y)\| \\
\leq&\eta\frac{cd_2}{\delta}\|\nabla_y g(x_1,y)-\eta\nabla_y g(x_2,y)\| \\
\leq&\eta\frac{cd_2}{\delta}L_g\|x_1-x_2\|
\end{aligned} \tag{46}$$

$$\begin{aligned}
&\|\left[I_{d_2}-(I_{d_2}-\eta\nabla^2_{yy}g(x_2,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)^\top\right]^{-1}-\left[I_{d_2}-(I_{d_2}-\eta\nabla^2_{yy}g(x_1,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)^\top\right]^{-1}\| \\
\leq&\frac{1}{\eta^2\mu_g^2}\|(I_{d_2}-\eta\nabla^2_{yy}g(x_1,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)^\top-(I_{d_2}-\eta\nabla^2_{yy}g(x_2,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)^\top\| \\
\leq&\frac{1}{\eta^2\mu_g^2}\|\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_1)-\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)\|\|I_{d_2}-\eta\nabla^2_{yy}g(x_1,y)\| \\
&+\frac{1}{\eta^2\mu_g^2}\|\nabla\mathcal{P}_{\mathcal{Y}\delta}(z_2)\|\|I_{d_2}-\eta\nabla^2_{yy}g(x_1,y)-I_{d_2}+\eta\nabla^2_{yy}g(x_2,y)\| \\
\leq&\left(\frac{1}{\eta\mu_g^2}\frac{cd_2}{\delta}L_g(1-\eta\mu_g)+\frac{L_{gyy}}{\eta\mu_g^2}\right)\|x_1-x_2\|
\end{aligned} \tag{47}$$

Therefore, we have

$$\begin{aligned}&\|\nabla f_\delta(x_1,y)-\nabla f_\delta(x_2,y)\|\\ \le&\left(L_f+\frac{L_{gxy}C_{fy}}{\mu_g}+\frac{C_{gxy}C_{fy}}{\mu_g}\eta\frac{cd_2}{\delta}L_g+C_{gxy}C_{fy}\left(\eta\frac{1}{\mu_g^2}\frac{cd_2}{\delta}L_g(1-\eta\mu_g)+\frac{L_{gyy}}{\mu_g^2}\right)+\frac{C_{gxy}}{\mu_g}L_f\right)\|x_1-x_2\|\end{aligned}\tag{48}$$

□

G.1. Proof of Lemma 3.8

Here we first give a detailed version of Lemma 3.8 and then give the proof.

Lemma G.2. ***(Lipschitz continous of $\nabla F_\delta(x)$.)*** *Under Assumptions 2.1, 2.2 and Lemma 3.10, we have $\nabla F_\delta(x)$ is Lipschitz continuous w.r.t x, such that*

$$\|\nabla F_\delta(x_1)-\nabla F_\delta(x_2)\|\le L_{F_\delta}(\frac{d_2}{\delta})\|x_1-x_2\|\tag{49}$$

where $L_{F_\delta}(\frac{d_2}{\delta})=\frac{L_g}{\mu_g}L_1(\frac{d_2}{\delta})+L_2(\frac{d_2}{\delta})$, $L_1(\frac{d_2}{\delta})=L_f+\frac{L_{gxy}C_{fy}}{\mu_g}+\frac{C_{gxy}C_{fy}}{\mu_g}(\frac{cd_2}{\delta}+\eta\frac{cd_2}{\delta}L_g)+C_{gxy}C_{fy}(\frac{1}{\eta\mu_g^2}(\frac{cd_2}{\delta}+\eta\frac{cd_2}{\delta}L_g)(1-\eta\mu_g)+\frac{L_{gyy}}{\mu_g^2})+\frac{C_{gxy}}{\mu_g}L_f$ *and* $L_2(\frac{d_2}{\delta})=L_f+\frac{L_{gxy}C_{fy}}{\mu_g}+\frac{C_{gxy}C_{fy}}{\mu_g}\eta\frac{cd_2}{\delta}L_g+C_{gxy}C_{fy}(\eta\frac{1}{\mu_g^2}\frac{cd_2}{\delta}L_g(1-\eta\mu_g)+\frac{L_{gyy}}{\mu_g^2})+\frac{C_{gxy}}{\mu_g}L_f$.

Proof. We have

$$\begin{aligned}&\|\nabla F_\delta(x_1)-\nabla F_\delta(x_2)\|\\ =&\|\nabla f_\delta(x_1,y^*(x_1))-\nabla f_\delta(x_2,y^*(x_2))\|\\ \le&\|\nabla f_\delta(x_1,y^*(x_1))-\nabla f_\delta(x_1,y^*(x_2))\|+\|f_\delta(x_1,y^*(x_2))-f_\delta(x_2,y^*(x_2))\|\end{aligned}\tag{50}$$

For the first term, we have

$$\|\nabla f_\delta(x_1,y^*(x_1))-\nabla f_\delta(x_1,y^*(x_2))\|\le L_1(\frac{d_2}{\delta})\|y^*(x_1)-y^*(x_2)\|\le\frac{L_g}{\mu_g}L_1(\frac{d_2}{\delta})\|x_1-x_2\|\tag{51}$$

Thus, we have

$$\|\nabla F_\delta(x_1)-\nabla F_\delta(x_2)\|\le\left(L_2(\frac{d_2}{\delta})+\frac{L_g}{\mu_g}L_1(\frac{d_2}{\delta})\right)\|x_1-x_2\|\tag{52}$$

□

H. Proof of Lemma 3.10

Proof. By the definition of $\bar{H}$, we have

$$\bar{H}=\sum_{i=1}^{d_2}\frac{1}{2\delta}\left(\mathcal{P}_{\mathcal{Y}}(z+\delta u_i)-\mathcal{P}_{\mathcal{Y}}(z-\delta u_i)\right)u_i^\top=\begin{bmatrix}\sum_{i=1}^{d_2}\frac{1}{2\delta}\left(\mathcal{P}_{\mathcal{Y}}^1(z+\delta u_i)-\mathcal{P}_{\mathcal{Y}}^1(z-\delta u_i)\right)u_i^\top\\ \vdots\\ \sum_{i=1}^{d_2}\frac{1}{2\delta}\left(\mathcal{P}_{\mathcal{Y}}^{d_2}(z+\delta u_i)-\mathcal{P}_{\mathcal{Y}}^{d_2}(z-\delta u_i)\right)u_i^\top\end{bmatrix}\tag{53}$$

According to (Lin et al., 2022), for each element $\bar{H}_j=\sum_{i=1}^{d_2}\frac{1}{2\delta}\left(\mathcal{P}_{\mathcal{Y}}^j(z+\delta u_i)-\mathcal{P}_{\mathcal{Y}}^j(z-\delta u_i)\right)u_i^\top$, we have

$$\mathbb{E}[\frac{1}{d_2}\sum_{i=1}^{d_2}\frac{d_2}{2\delta}\left(\mathcal{P}_{\mathcal{Y}}^j(z+\delta u_i)-\mathcal{P}_{\mathcal{Y}}^j(z-\delta u_i)\right)u_i^\top]=\nabla\mathcal{P}_{\mathcal{Y}\delta}^j(z).\tag{54}$$

Therefore, we have $\mathbb{E}[\bar{H}] = \nabla\mathcal{P}_{\mathcal{Y}\delta}(z)$. According to (Lin et al., 2022), we have $\mathbb{E}_{u_i}[\|\frac{d_2}{2\delta}\left(\mathcal{P}^j_{\mathcal{Y}}(z+\delta u_i) - \mathcal{P}^j_{\mathcal{Y}}(z-\delta u_i)\right)u_i^\top\|^2] \leq 16\sqrt{2\pi}d_2L_p^2$. Then, we have

$$
\begin{aligned}
&\mathbb{E}[\|\bar{H}\|^2] \\
\leq&\mathbb{E}[\|\bar{H}\|_F^2] = \mathbb{E}[\sum_{j=1}^{d_2}\|\bar{H}_j\|^2] = \sum_{j=1}^{d_2}\mathbb{E}[\|\bar{H}_j\|^2] \\
=&\sum_{j=1}^{d_2}\mathbb{E}\left[\left\|\frac{1}{d_2}\sum_{i=1}^{d_2}\frac{d_2}{2\delta}\left(\mathcal{P}^j_{\mathcal{Y}}(z+\delta u_i) - \mathcal{P}^j_{\mathcal{Y}}(z-\delta u_i)\right)u_i^\top\right\|^2\right] \\
\leq&\sum_{j=1}^{d_2}\frac{1}{d_2^2}\sum_{i=1}^{d_2}\mathbb{E}_{u_i}\left[\left\|\frac{d_2}{2\delta}\left(\mathcal{P}^j_{\mathcal{Y}}(z+\delta u_i) - \mathcal{P}^j_{\mathcal{Y}}(z-\delta u_i)\right)u_i^\top\right\|^2\right] \\
\leq&d_2\cdot\frac{1}{d_2^2}d_2\cdot 16\sqrt{2\pi}d_2L_p^2 \\
=&16\sqrt{2\pi}d_2L_p^2
\end{aligned}
\tag{55}
$$

That completes the proof. □

I. Proof of Lemma 3.11

Proof. For convenience, define $\bar{G}_{yy} = Q\prod_{i=1}^{c(Q)}\left((I_{d_2} - \eta\nabla^2_{yy}g(x,y))\bar{H}(z;u^i)^\top\right)$ and $G_{yy} = \left[I_{d_2} - (I_{d_2} - \eta\nabla^2_{yy}g(x,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)^\top\right]^{-1}$. We set $\eta < \frac{1}{\mu_g}$ We have

$$
\mathbb{E}_{\bar{\xi}}[\bar{\nabla}f_\delta(x,y;\bar{\xi})] = \nabla_x f(x,y) - \eta\nabla^2_{xy}g(x,y)\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)^\top\mathbb{E}\left[\bar{G}_{yy}\right]\nabla_y f(x,y) \tag{56}
$$

We have

$$
\begin{aligned}
&\left\|\nabla f_\delta(x,y) - \mathbb{E}[\bar{\nabla}f_\delta(x,y;\bar{\xi})]\right\| \\
=&\left\|\eta\nabla^2_{xy}g(x,y)\partial_z\mathcal{P}_{\mathcal{Y}\delta}(z)^\top\left\{\mathbb{E}\left[\bar{G}_{yy}\right] - G_{yy}\right\}\nabla_y f(x,y)\right\| \\
\leq&\eta C_{gxy}C_{fy}\left\|\mathbb{E}\left[\bar{G}_{yy}\right] - G_{yy}\right\|
\end{aligned}
\tag{57}
$$

where the third inequality is due to the non-expansive of the projector operation.

Due to the independency of $u, c(Q)$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\bar{G}_{yy}\right] \\
=&\mathbb{E}\left[Q\prod_{i=1}^{c(Q)}\left((I_{d_2} - \eta\nabla^2_{yy}g(x,y))\bar{H}(z;u^i)^\top\right)\right] \\
=&Q\mathbb{E}_{c(Q)}\left[\mathbb{E}_u\left[\prod_{i=1}^{c(Q)}\left((I_{d_2} - \eta\nabla^2_{yy}g(x,y))\bar{H}(z;u^i)^\top\right)\right]\right] \\
=&Q\mathbb{E}_{c(Q)}\left[(I_{d_2} - \eta\nabla^2_{yy}g(x,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)^\top\right]^{c(Q)} \\
=&\sum_{i=0}^{Q-1}\left[(I_{d_2} - \eta\nabla^2_{yy}g(x,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)^\top\right]^i
\end{aligned}
\tag{58}
$$

In addition, we have

$$
\begin{aligned}
&G_{yy}\\
=&\left[I_{d_2}-(I_{d_2}-\eta\nabla^2_{yy}g(x,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)^\top\right]^{-1}\\
=&\sum_{i=0}^{\infty}\left[(I_{d_2}-\eta\nabla^2_{yy}g(x,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)^\top\right]^i\\
=&\mathbb{E}\left[Q\prod_{i=1}^{c(Q)}\left((I_{d_2}-\eta\nabla^2_{yy}g(x,y))\bar{H}(z;u^i)^\top\right)\right]+\sum_{i=Q}^{\infty}\left[(I_{d_2}-\eta\nabla^2_{yy}g(x,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)^\top\right]^i
\end{aligned}
\tag{59}
$$

where $z=y-\eta\nabla_y g(x,y)$, which implies that

$$
\begin{aligned}
&\left\|\mathbb{E}\left[\bar{G}_{yy}\right]-\bar{G}_{yy}\right\|\\
=&\left\|\left[(I_{d_2}-(I_{d_2}-\eta\nabla^2_{yy}g(x,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)^\top\right]^{-1}-\mathbb{E}\left[Q\prod_{i=1}^{c(Q)}\left((I_{d_2}-\eta\nabla^2_{yy}g(x,y;\zeta^i))\bar{H}(z;u^i)^\top\right)\right]\right\|\\
\leq&\sum_{i=Q}^{\infty}\left\|(I_{d_2}-\eta\nabla^2_{yy}g(x,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)^\top\right\|^i\\
\leq&\sum_{i=Q}^{\infty}\left\|I_{d_2}-\eta\nabla^2_{yy}g(x,y))\right\|^i\left\|\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)\right\|^i\\
\leq&\frac{1}{\eta\mu_g}(1-\eta\mu_g)^Q
\end{aligned}
\tag{60}
$$

Thus, we have

$$
\left\|\nabla f_\delta(x,y)-\mathbb{E}[\bar{\nabla}f_\delta(x,y;\bar{\xi})]\right\|\leq\frac{C_{gxy}C_{fy}}{\mu_g}(1-\eta\mu_g)^Q
\tag{61}
$$

Then, we prove the bound on the variance. we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\bar{\nabla}f_\delta(x,y;\bar{\xi})-\mathbb{E}\left[\bar{\nabla}f_\delta(x,y;\bar{\xi})\right]\right\|^2\right]\\
=&\mathbb{E}\left[\left\|\eta\nabla^2_{xy}g(x,y)\bar{H}(z;u^0)^\top\bar{G}_{yy}\nabla_y f(x,y)-\eta\nabla^2_{xy}g(x,y)\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)^\top\mathbb{E}[\bar{G}_{yy}]\nabla_y f(x,y)\right\|^2\right]\\
\leq&2\eta^2\left\|\nabla^2_{xy}g(x,y)\right\|^2\mathbb{E}\left[\left\|\bar{H}(z;u^0)-\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)\right\|^2\right]\mathbb{E}\left[\left\|\bar{G}_{yy}\right\|^2\right]\left\|\nabla_y f(x,y)\right\|^2\\
&+2\eta^2\left\|\nabla_{xy}g(x,y)\right\|^2\left\|\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)\right\|^2\mathbb{E}\left[\left\|\bar{G}_{yy}-\mathbb{E}[\bar{G}_{yy}]\right\|^2\right]\left\|\nabla_y f(x,y)\right\|^2
\end{aligned}
\tag{62}
$$

For the first term in the above inequality, we have

$$
\mathbb{E}\left[\left\|\bar{H}(z;u^0)-\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)\right\|^2\right]\leq 32\sqrt{2\pi}d_2L_p^2+2
\tag{63}
$$

For $\mathbb{E}[\|\bar{G}_{yy}\|^2]$, we have

$$\begin{aligned}\mathbb{E}[\|\bar{G}_{yy}\|^2] &= \sum_{q=0}^{Q-1}\mathbb{E}\left[\left\|\prod_{i=1}^{q}\left((I_{d_2}-\eta\nabla^2_{yy}g(x,y))\bar{H}(z;u^i)^\top\right)\right\|^2\right]\\ &\leq \sum_{q=0}^{Q-1}((1-\eta\mu_g)4(2\pi)^{1/4}\sqrt{d_2}L_p)^{2q}\\ &\leq \frac{1}{1-(1-\eta\mu_g)^2 16\sqrt{2\pi}d_2L_p^2}\end{aligned} \tag{64}$$

where the last inequality is obtained by setting $\frac{1}{\mu_g}(1-\frac{1}{4(2\pi)^{1/4}\sqrt{d_2}L_p})\leq\eta<\frac{1}{\mu_g}$.

We can also derive that

$$\begin{aligned}&\left\|\mathbb{E}\left[\bar{G}_{yy}\right]\right\|\\ =&\|\sum_{i=0}^{Q-1}\left[(I_{d_2}-\eta\nabla^2_{yy}g(x,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)^\top\right]^i\|\\ \leq&\sum_{i=0}^{Q-1}\|(I_{d_2}-\eta\nabla^2_{yy}g(x,y))\nabla\mathcal{P}_{\mathcal{Y}\delta}(z)^\top\|^i\\ \leq&\sum_{i=0}^{Q-1}(1-\eta\mu_g)^i\\ \leq&\frac{1}{\eta\mu_g}\end{aligned} \tag{65}$$

Then, we have

$$\mathbb{E}\left[\|G_{yy}-\mathbb{E}[G_{yy}]\|^2\right]\leq\frac{2}{1-(1-\eta\mu_g)^2 16\sqrt{2\pi}d_2L_p^2}+\frac{2}{\eta^2\mu_g^2} \tag{66}$$

Therefore, combining the above inequalities, we can bound the variance as follows,

$$\mathbb{E}\left[\left\|\bar{\nabla}f(x,y;\bar{\xi})-\mathbb{E}\left[\bar{\nabla}f(x,y;\bar{\xi})\right]\right\|^2\right]\leq\sigma_f(d_2) \tag{67}$$

where $\sigma_f(d_2)=2\eta^2C_{gxy}^2(32\sqrt{2\pi}d_2L_p^2+2)C_{fy}^2\frac{1}{1-(1-\eta\mu_g)^2 16\sqrt{2\pi}d_2L_p^2}+2\eta^2C_{gxy}^2(\frac{2}{1-(1-\eta\mu_g)^2 16\sqrt{2\pi}d_2L_p^2}+\frac{2}{\eta^2\mu_g^2})C_{fy}^2$ That completes the proof. □

J. Route Map of Our Convergence Analysis

Here we give a simple route map of our convergence analysis.

K. Deriving the Convergence Metric

In this section, we give a detailed analysis to derive the metric. Based on the Lipshitz continuous of $\nabla F_\delta(x)$, we have the following lemma,

Lemma K.1. *Under Assumptions 2.1, 2.2 and Lemma 3.10, we have*

$$\|\nabla F_\delta(x_k)-w_k\|^2\leq 2L_1^2(\frac{d_2}{\delta})\|y^*(x_k)-y_k\|^2+2\|\nabla f_\delta(x_k,y_k)-w_k\|^2. \tag{68}$$

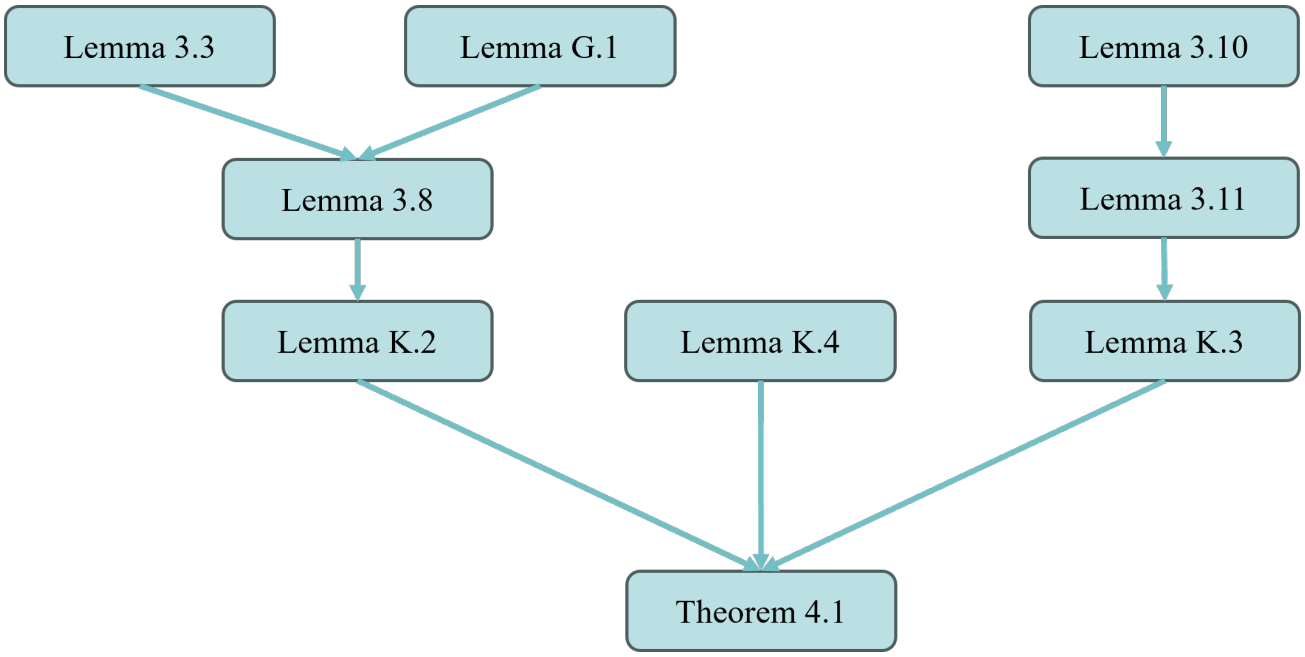


Figure 3: Route map of convergence analysis.

Proof. We have

$$\begin{aligned}
&\|\nabla F_\delta(x_k) - w_k\|^2 \\
=&\|\nabla f_\delta(x_k, y^*(x_k)) - \nabla f_\delta(x_k, y_k) + \nabla f_\delta(x_k, y_k) - w_k\|^2 \\
\leq& 2\|\nabla f_\delta(x_k, y^*(x_k)) - \nabla f_\delta(x_k, y_k)\|^2 + 2\|\nabla f_\delta(x_k, y_k) - w_k\|^2 \\
\leq& 2L_1^2(\frac{d_2}{\delta})\|y^*(x_k) - y_k\|^2 + 2\|\nabla f_\delta(x_k, y_k) - w_k\|^2
\end{aligned} \tag{69}$$

□

K.1. Usefull Lemmas in Convergence Rate

Lemma K.2. ***(Descent on the function value.)*** *Under Assumptions 2.1, 2.2, and Lemma 3.8, let $F_\delta(x)$ be an approximation function of $F(x)$ and have the gradient $\nabla F_\delta(x_k)$, and $\gamma\eta_k \leq \frac{1}{2L_{F_\delta}(\frac{d_2}{\delta})c_l}$, we have*

$$F_\delta(x_{k+1}) \leq F_\delta(x_k) + \eta_k\gamma c_l\|\nabla F_\delta(x_k) - w_k\|^2 - \frac{\eta_k}{2\gamma c_l}\|\tilde{x}_{k+1} - x_k\|^2 \tag{70}$$

Proof. Due to the smoothness of F_δ and let $\tilde{x}_{k+1} = x_k - \frac{\gamma}{\mathcal{P}_{[1/c_u,1/c_l]}(\sqrt{m_{2,k}}+G_0)}w_k$, we have

$$\begin{aligned}
&F_\delta(x_{k+1}) \\
\leq& F_\delta(x_k) + \nabla F_\delta(x_k)^\top(x_{k+1} - x_k) + \frac{1}{2}L_{F_\delta}(\frac{d_2}{\delta})\|x_{k+1} - x_k\|^2 \\
=& F_\delta(x_k) + \eta_k\nabla F_\delta(x_k)^\top(\tilde{x}_{k+1} - x_k) + \frac{1}{2}L_{F_\delta}(\frac{d_2}{\delta})\|\eta_k(\tilde{x}_{k+1} - x_k)\|^2 \\
=& F_\delta(x_k) + \eta_k\langle w_k, \tilde{x}_{k+1} - x_k\rangle + \eta_k\langle\nabla F_\delta(x_k) - w_k, \tilde{x}_{k+1} - x_k\rangle + \frac{\eta_k^2}{2}L_{F_\delta}(\frac{d_2}{\delta})\|\tilde{x}_{k+1} - x_k\|^2
\end{aligned} \tag{71}$$

In our algorithm, we have $\tilde{x}_{k+1} = x_k - \frac{\gamma}{\mathcal{P}_{[1/c_u,1/c_l]}(\sqrt{m_{2,k}}+G_0)}w_k = \arg\min_{x\in\mathbb{R}^{d_1}} \frac{1}{2}\|x - x_k + \frac{\gamma}{\mathcal{P}_{[1/c_u,1/c_l]}(\sqrt{m_{2,k}}+G_0)}w_k\|^2$.

We have the following optimal condition,

$$\langle \tilde{x}_{k+1} - x_k + \frac{\gamma}{\mathcal{P}_{[1/c_u,1/c_l]}(\sqrt{m_{2,k}} + G_0)} w_k, x - \tilde{x}_{k+1} \rangle \geq 0,\ x \in \mathbb{R}^{d_1} \tag{72}$$

Set $x = x_k$, we can obtain

$$\gamma c_l \langle w_k, \tilde{x}_{k+1} - x_k \rangle \leq -\|\tilde{x}_{k+1} - x_k\|^2 \tag{73}$$

Thus, we have

$$\langle w_k, \tilde{x}_{k+1} - x_k \rangle \leq -\frac{1}{\gamma c_l} \|\tilde{x}_{k+1} - x_k\|^2 \tag{74}$$

In addition, we can obtain

$$\begin{aligned} &\langle \nabla F_\delta(x_k) - w_k, \tilde{x}_{k+1} - x_k \rangle \\ \leq &\|\nabla F_\delta(x_k) - w_k\|_2 \|\tilde{x}_{k+1} - x_k\|_2 \\ \leq &\gamma c_l \|\nabla F_\delta(x_k) - w_k\|^2 + \frac{1}{4\gamma c_l} \|\tilde{x}_{k+1} - x_k\|^2 \end{aligned} \tag{75}$$

Then, setting $\gamma \leq \frac{1}{2L_{F_\delta}(\frac{d_2}{\delta}) c_l \eta_k}$, we can derive

$$\begin{aligned} &F_\delta(x_{k+1}) \\ \leq &F_\delta(x_k) + \eta_k \gamma c_l \|\nabla F_\delta(x_k) - w_k\|^2 + \frac{\eta_k}{4\gamma c_l} \|\tilde{x}_{k+1} - x_k\|^2 - \frac{\eta_k}{\gamma c_l} \|\tilde{x}_{k+1} - x_k\|^2 + \frac{\eta_k^2}{2} L_{F_\delta}(\frac{d_2}{\delta}) \|\tilde{x}_{k+1} - x_k\|^2 \\ \leq &F_\delta(x_k) + \eta_k \gamma c_l \|\nabla F_\delta(x_k) - w_k\|^2 - \frac{\eta_k}{2\gamma c_l} \|\tilde{x}_{k+1} - x_k\|^2 \end{aligned} \tag{76}$$

□

Lemma K.3. ***(Error between the updates of y and the optimal solution y^*)*** *Under Assumptions 2.1, 2.2, let $\tilde{y}_{k+1} = \mathcal{P}_{\mathcal{Y}}(y_k - \frac{\tau}{\mathcal{P}_{[1/c_u,1/c_l]}(\sqrt{m_{1,k}}+G_0)} v_k)$ and $\eta_k \leq 1$, $\tau \leq \frac{1}{6L_g c_u}$, we have*

$$\begin{aligned} &\|y_{k+1} - y^*(x_{k+1})\|^2 \\ \leq &(1 - \frac{\eta_k \tau \mu_g c_u}{4}) \|y^*(x_k) - y_k\|^2 + \frac{25\eta_k \tau c_u}{6\mu_g} \|\nabla_y g(x_k, y_k) - v_k\|^2 - \frac{3\eta_k}{4} \|\tilde{y}_{k+1} - y_k\|^2 + \frac{25L_y^2 \eta_k}{6\tau \mu_g c_u} \|x_k - \tilde{x}_{k+1}\|^2 \end{aligned} \tag{77}$$

Proof. Define $\tilde{y}_{k+1} = \mathcal{P}_{\mathcal{Y}}(y_k - \frac{\tau}{\mathcal{P}_{[1/c_u,1/c_l]}(\sqrt{m_{1,k}}+G_0)} v_k)$. We have $y_{k+1} = (1 - \eta_k) y_k + \eta_k \tilde{y}_{k+1}$.

According to the strong convexity of g, we have

$$\begin{aligned} &g(x_k, y) \\ \geq &g(x_k, y_k) + \langle \nabla_y g(x_k, y_k), y - y_k \rangle + \frac{\mu_g}{2} \|y - y_k\|^2 \\ = &g(x_k, y_k) + \langle v_k, y - \tilde{y}_{k+1} \rangle + \langle \nabla_y g(x_k, y_k) - v_k, y - \tilde{y}_{k+1} \rangle + \langle \nabla_y g(x_k, y_k), \tilde{y}_{k+1} - y_k \rangle + \frac{\mu_g}{2} \|y - y_k\|^2 \end{aligned} \tag{78}$$

According to the smoothness of g, we have

$$g(x_k, \tilde{y}_{k+1}) \leq g(x_k, y_k) + \langle \nabla_y g(x_k, y_k), \tilde{y}_{k+1} - y_k \rangle + \frac{L_g}{2} \|\tilde{y}_{k+1} - y_k\|^2 \tag{79}$$

Then, combining the above inequalities, we have

$$g(x_k, y) \geq g(x_k, \tilde{y}_{k+1}) - \frac{L_g}{2} \|\tilde{y}_{k+1} - y_k\|^2 + \langle v_k, y - \tilde{y}_{k+1} \rangle + \langle \nabla_y g(x_k, y_k) - v_k, y - \tilde{y}_{k+1} \rangle + \frac{\mu_g}{2} \|y - y_k\|^2 \tag{80}$$

In our Algorithm 1, we have

$$\tilde{y}_{k+1} = \mathcal{P}_{\mathcal{Y}}\left(y_k - \frac{\tau}{\mathcal{P}_{[1/c_u,1/c_l]}(\sqrt{m_{1,k}}+G_0)}v_k\right) = \arg\min_{y\in\mathcal{Y}} \frac{1}{2}\left\|y - y_k + \frac{\tau}{\mathcal{P}_{[1/c_u,1/c_l]}(\sqrt{m_{1,k}}+G_0)}v_k\right\|^2. \tag{81}$$

Since $\mathcal{Y}$ is a convex set and the function $\frac{1}{2}\left\|y - y_k + \frac{\tau}{\mathcal{P}_{[1/c_u,1/c_l]}(\sqrt{\|v_k\|})}v_k\right\|^2$ is convex, we have

$$\left\langle \tilde{y}_{k+1} - y_k + \frac{\tau}{\mathcal{P}_{[1/c_u,1/c_l]}(\sqrt{m_{1,k}}+G_0)}v_k, y - \tilde{y}_{k+1}\right\rangle \geq 0, \quad y \in \mathcal{Y} \tag{82}$$

Then we have

$$\tau c_u \langle v_k, y - \tilde{y}_{k+1}\rangle \geq \langle \tilde{y}_{k+1} - y_k, \tilde{y}_{k+1} - y\rangle \tag{83}$$

Then we have

$$\begin{aligned} & g(x_k, y) \\ \geq & g(x_k, \tilde{y}_{k+1}) - \frac{L_g}{2}\|\tilde{y}_{k+1} - y_k\|^2 + \frac{\mu_g}{2}\|y - y_k\|^2 + \langle \nabla_y g(x_k, y_k) - v_k, y - \tilde{y}_{k+1}\rangle + \frac{1}{\tau c_u}\langle \tilde{y}_{k+1} - y_k, \tilde{y}_{k+1} - y\rangle \end{aligned} \tag{84}$$

Let $y = y^*(x_k)$. Since $g(x_k, y^*(x_k)) \leq g(x_k, \tilde{y}_{k+1})$, we have

$$\begin{aligned} & g(x_k, \tilde{y}_{k+1}) \geq g(x_k, y^*(x_k)) \\ \geq & g(x_k, \tilde{y}_{k+1}) - \frac{L_g}{2}\|\tilde{y}_{k+1} - y_k\|^2 + \frac{\mu_g}{2}\|y^*(x_k) - y_k\|^2 + \langle \nabla_y g(x_k, y_k) - v_k, y^*(x_k) - \tilde{y}_{k+1}\rangle \\ & + \frac{1}{\tau c_u}\|\tilde{y}_{k+1} - y_k\|^2 + \frac{1}{\tau c_u}\langle \tilde{y}_{k+1} - y_k, y_k - y^*(x_k)\rangle \end{aligned} \tag{85}$$

In addition, we have

$$\begin{aligned} & \langle \nabla_y g(x_k, y_k) - v_k, y^*(x_k) - \tilde{y}_{k+1}\rangle \\ = & \langle \nabla_y g(x_k, y_k) - v_k, y^*(x_k) - y_k\rangle + \langle \nabla_y g(x_k, y_k) - v_k, y_k - \tilde{y}_{k+1}\rangle \\ \geq & -\|\nabla_y g(x_k, y_k) - v_k\|\|y^*(x_k) - y_k\| - \|\nabla_y g(x_k, y_k) - v_k\|\|y_k - \tilde{y}_{k+1}\| \\ \geq & -\frac{1}{\mu_g}\|\nabla_y g(x_k, y_k) - v_k\|^2 - \frac{\mu_g}{4}\|y^*(x_k) - y_k\|^2 - \frac{1}{\mu_g}\|\nabla_y g(x_k, y_k) - v_k\|^2 - \frac{\mu_g}{4}\|y_k - \tilde{y}_{k+1}\|^2 \\ \geq & -\frac{2}{\mu_g}\|\nabla_y g(x_k, y_k) - v_k\|^2 - \frac{\mu_g}{4}\|y^*(x_k) - y_k\|^2 - \frac{\mu_g}{4}\|y_k - \tilde{y}_{k+1}\|^2 \end{aligned} \tag{86}$$

The first inequality is due to $\langle a, b\rangle \geq -\|a\|\|b\|$ and the second inequality is due to the Young's inequality. We also have

$$\begin{aligned} & \|y_{k+1} - y^*(x_k)\|^2 \\ \leq & \|y_k + \eta_k(\tilde{y}_{k+1} - y_k) - y^*(x_k)\|^2 \\ = & \|y_k - y^*(x_k)\|^2 + \eta_k^2\|\tilde{y}_{k+1} - y_k\|^2 + 2\eta_k\langle \tilde{y}_{k+1} - y_k, y_k - y^*(x_k)\rangle \end{aligned} \tag{87}$$

Therefore, we have

$$\langle \tilde{y}_{k+1} - y_k, y_k - y^*(x_k)\rangle \geq \frac{1}{2\eta_k}(\|y_{k+1} - y^*(x_k)\|^2 - \|y_k - y^*(x_k)\|^2 - \eta_k^2\|\tilde{y}_{k+1} - y_k\|^2) \tag{88}$$

Then, we have

$$\begin{aligned} 0 \geq & -\frac{L_g}{2}\|\tilde{y}_{k+1} - y_k\|^2 + \frac{\mu_g}{2}\|y^*(x_k) - y_k\|^2 + \frac{1}{\tau c_u}\|\tilde{y}_{k+1} - y_k\|^2 \\ & - \frac{2}{\mu_g}\|\nabla_y g(x_k, y_k) - v_k\|^2 - \frac{\mu_g}{4}\|y^*(x_k) - y_k\|^2 - \frac{\mu_g}{4}\|y_k - \tilde{y}_{k+1}\|^2 \\ & + \frac{1}{2\eta_k\tau c_u}(\|y_{k+1} - y^*(x_k)\|^2 - \|y_k - y^*(x_k)\|^2 - \eta_k^2\|\tilde{y}_{k+1} - y_k\|^2) \end{aligned} \tag{89}$$

Hence we have

$$\begin{aligned}
&\|y_{k+1}-y^*(x_k)\|^2\\
\leq&2\eta_k\tau c_u(\frac{L_g}{2}-\frac{1}{\tau c_u}+\frac{\mu_g}{4}+\frac{\eta_k}{2\tau c_u})\|\tilde{y}_{k+1}-y_k\|^2+(1-\frac{\mu_g\eta_k\tau c_u}{2})\|y^*(x_k)-y_k\|^2+\frac{4\eta_k\tau c_u}{\mu_g}\|\nabla_y g(x_k,y_k)-v_k\|^2\\
\leq&(1-\frac{\mu_g\eta_k\tau c_u}{2})\|y^*(x_k)-y_k\|^2+\frac{4\eta_k\tau c_u}{\mu_g}\|\nabla_y g(x_k,y_k)-v_k\|^2-2\eta_k\tau c_u(\frac{1}{2\tau c_u}-\frac{3L_g}{4})\|\tilde{y}_{k+1}-y_k\|^2\\
\leq&(1-\frac{\mu_g\eta_k\tau c_u}{2})\|y^*(x_k)-y_k\|^2+\frac{4\eta_k\tau c_u}{\mu_g}\|\nabla_y g(x_k,y_k)-v_k\|^2-\frac{3\eta_k}{4}\|\tilde{y}_{k+1}-y_k\|^2
\end{aligned}\tag{90}$$

using $\eta_k\leq 1$, $\mu_g\leq L_g$ and $\tau\leq\frac{1}{6L_gc_u}$.

Then, we have

$$\begin{aligned}
&\|y_{k+1}-y^*(x_{k+1})\|^2\\
=&\|y_{k+1}-y^*(x_k)+y^*(x_k)-y^*(x_{k+1})\|^2\\
\leq&(1+\frac{\eta_k\tau\mu_gc_u}{4})\|y_{k+1}-y^*(x_k)\|^2+(1+\frac{4}{\eta_k\tau\mu_gc_u})\|y^*(x_k)-y^*(x_{k+1})\|^2\\
\leq&(1+\frac{\eta_k\tau\mu_gc_u}{4})\|y_{k+1}-y^*(x_k)\|^2+(1+\frac{4}{\eta_k\tau\mu_gc_u})L_y^2\|x_k-x_{k+1}\|^2\\
\leq&(1+\frac{\eta_k\tau\mu_gc_u}{4})(1-\frac{\mu_g\eta_k\tau c_u}{2})\|y^*(x_k)-y_k\|^2+(1+\frac{\eta_k\tau\mu_gc_u}{4})\frac{4\eta_k\tau c_u}{\mu_g}\|\nabla_y g(x_k,y_k)-v_k\|^2\\
&-(1+\frac{\eta_k\tau\mu_gc_u}{4})\frac{3\eta_k}{4}\|\tilde{y}_{k+1}-y_k\|^2+(1+\frac{4}{\eta_k\tau\mu_gc_u})L_y^2\|x_k-x_{k+1}\|^2
\end{aligned}\tag{91}$$

Since $\eta_k\leq 1$, $\mu_g\leq L_g$ and $\tau\leq\frac{1}{6L_gc_u}$, we have $\tau\leq\frac{1}{6L_gc_u}\leq\frac{1}{6\mu_gc_u}$ and $\eta_k\leq 1\leq\frac{1}{6\tau L_gc_u}$. Then, we can obtain

$$(1+\frac{\eta_k\tau\mu_gc_u}{4})(1-\frac{\mu_g\eta_k\tau c_u}{2})=1-\frac{\mu_g\eta_k\tau c_u}{2}+\frac{\eta_k\tau\mu_gc_u}{4}-\frac{\eta_k^2\tau^2\mu_g^2c_u^2}{8}\leq 1-\frac{\eta_k\tau\mu_gc_u}{4}\tag{92}$$

$$-(1+\frac{\eta_k\tau\mu_gc_u}{4})\frac{3\eta_k}{4}\leq-\frac{3\eta_k}{4}\tag{93}$$

$$(1+\frac{\eta_k\tau\mu_gc_u}{4})\frac{4\eta_k\tau c_u}{\mu_g}\leq\frac{25\eta_k\tau c_u}{6\mu_g}\tag{94}$$

$$(1+\frac{4}{\eta_k\tau\mu_gc_u})L_y^2\leq\frac{25L_y^2}{6\eta_k\tau\mu_gc_u}\tag{95}$$

Finally, we can obtain

$$\begin{aligned}
&\|y_{k+1}-y^*(x_{k+1})\|^2\\
\leq&(1-\frac{\eta_k\tau\mu_gc_u}{4})\|y^*(x_k)-y_k\|^2+\frac{25\eta_k\tau c_u}{6\mu_g}\|\nabla_y g(x_k,y_k)-v_k\|^2\\
&-\frac{3\eta_k}{4}\|\tilde{y}_{k+1}-y_k\|^2+\frac{25L_y^2}{6\eta_k\tau\mu_gc_u}\|x_k-x_{k+1}\|^2\\
\leq&(1-\frac{\eta_k\tau\mu_gc_u}{4})\|y^*(x_k)-y_k\|^2+\frac{25\eta_k\tau c_u}{6\mu_g}\|\nabla_y g(x_k,y_k)-v_k\|^2\\
&-\frac{3\eta_k}{4}\|\tilde{y}_{k+1}-y_k\|^2+\frac{25L_y^2\eta_k}{6\tau\mu_gc_u}\|x_k-\tilde{x}_{k+1}\|^2
\end{aligned}\tag{96}$$

□

Lemma K.4. ***(Descent in the gradient estimation error.(Huang & Huang, 2021))*** *Under Assumptions 2.1, 2.2, and Lemma*

3.11, if $\alpha \in (0,1)$ and $\beta \in (0,1)$, we have

$$\begin{aligned}&\mathbb{E}[\|\nabla f_\delta(x_{k+1},y_{k+1})+R_{k+1}-w_{k+1}\|^2]\\ \le&(1-\alpha)\mathbb{E}[\|\nabla f_\delta(x_k,y_k)+R_k-w_k\|^2]+\alpha^2\sigma_f(d_2)+\frac{3}{\alpha}(\|R_k\|^2+\|R_{k+1}\|^2)+\frac{3}{\alpha}L_0^2\eta_k^2(\|x_k-\tilde{x}_{k+1}\|^2+\|y_k-\tilde{y}_{k+1}\|^2)\end{aligned} \tag{97}$$

$$\begin{aligned}&\mathbb{E}[\|\nabla g(x_{k+1},y_{k+1})-v_{k+1}\|^2]\\ \le&(1-\beta)\mathbb{E}[\|\nabla g(x_k,y_k)-v_k\|^2]+\frac{2L_g^2}{\beta}\eta_k^2(\|x_k-\tilde{x}_{k+1}\|^2+\|y_k-\tilde{y}_{k+1}\|^2)\end{aligned} \tag{98}$$

where $L_0=\max(L_1(\frac{d_2}{\delta}),L_2(\frac{d_2}{\delta}))$.

K.2. Proof of the Convergence Rate in Theorem 4.1

Here we first give a detailed version of Theorem 4.1 and then present the proof.

Theorem K.5. *Under Assumptions 2.1, 2.2 and Lemma 3.10, with $\frac{1}{\mu_g}(1-\frac{1}{4(2\pi)^{1/4}\sqrt{d_2}L_p})\le\eta<\frac{1}{\mu_g}$, $Q=\frac{1}{\mu_g\eta}\ln\frac{C_{gxy}C_{fy}K}{\mu_g}$, $0\le a\le 2$, $0<\gamma\le\min\left\{\frac{1}{L_0^{2-a}},\frac{1}{2L_{F_\delta}(\frac{d_2}{\delta})c_l\eta_k},\frac{1}{4c_l\left(\frac{125L_0^aL_y^2c_l}{3\tau^2\mu_g^2c_u^2}+(\frac{2}{3}L_0^2+\frac{6\mu_g^2L_g^2}{125L_0^2})c_l\right)},\frac{2m^{1/2}}{9t}\right\}$, $0<\tau\le\min\left\{\frac{1}{6L_gc_u},\frac{15L_0^a}{2\mu_gc_u\left(\frac{2}{3}L_0^2+\frac{6\mu_g^2L_g^2}{125L_0^2}\right)}\right\}$, $m\ge\max\{t^2,c_1^2t^2,c_2^2t^2\}$, $\alpha=c_1\eta_k$, $\beta=c_2\eta_k$, $\frac{9}{2}\gamma\le c_1\le\frac{m^{1/2}}{t}$, $\frac{125L_0^2}{3\mu_g^2}\le c_2\le\frac{m^{1/2}}{t}$, $L_0=\max(L_1(\frac{d_2}{\delta}),L_2(\frac{d_2}{\delta}))>1$, $\Phi_1=\mathbb{E}[F_\delta(x_1)+\frac{10L_0^ac_l}{\tau\mu_gc_u}\|y_1-y^*(x_1)\|^2+c_l(\|w_1-\bar{\nabla}f_\delta(x_1,y_1)-R_1\|^2+\|\nabla_yg(x_1,y_1)-v_1\|^2)]$, and $\eta_k=\frac{t}{(m+k)^{1/2}}$, $t>0$, we have*

$$\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\frac{1}{2}\|\nabla F_\delta(x_k)\|]\le\frac{2m^{1/4}\sqrt{G}}{\sqrt{Kt}}+\frac{2\sqrt{G}}{(Kt)^{1/4}}. \tag{99}$$

where $G=\frac{\Phi_1-\Phi^*}{\gamma c_l}+\frac{17t}{4K^2}(m+K)^{1/2}+\frac{4}{3tK^2}(m+K)^{3/2}+(m\sigma_f(d_2))t^2\ln(m+K)$.

Proof. Setting $\eta_k=\frac{t}{(m+k)^{1/2}}$ and $m\ge t^2$, we have $\eta_k\le 1$. Due to $m\ge(c_1t)^2$, we have $\alpha=c_1\eta_k\le\frac{c_1t}{m^{1/2}}\le 1$. Due to $m\ge(c_2t)^2$, we have $\beta=c_2\eta_k\le\frac{c_2t}{m^{1/2}}\le 1$. Also, we have $c_1,c_2\le\frac{m^{1/2}}{t}$. Then using the above lemmas, we have

$$\begin{aligned}&\mathbb{E}[\|\nabla f_\delta(x_{k+1},y_{k+1})+R_{k+1}-w_{k+1}\|^2]-\mathbb{E}[\|\nabla f_\delta(x_k,y_k)+R_k-w_k\|^2]\\ \le&-\alpha\mathbb{E}[\|\nabla f_\delta(x_k,y_k)+R_k-w_k\|^2]+\alpha^2\sigma_f(d_2)+\frac{3}{\alpha}(\|R_k\|^2+\|R_{k+1}\|^2)\\ &+\frac{3}{\alpha}L_0^2\eta_k^2(\|x_k-\tilde{x}_{k+1}\|^2+\|y_k-\tilde{y}_{k+1}\|^2)\\ \le&-c_1\eta_k\mathbb{E}[\|\nabla f_\delta(x_k,y_k)+R_k-w_k\|^2]+c_1^2\eta_k^2\sigma_f(d_2)+\frac{3}{c_1\eta_k}(\|R_k\|^2+\|R_{k+1}\|^2)\\ &+\frac{3}{c_1}L_0^2\eta_k(\|x_k-\tilde{x}_{k+1}\|^2+\|y_k-\tilde{y}_{k+1}\|^2)\\ \le&-\frac{9}{2}\gamma\eta_k\mathbb{E}[\|\nabla f_\delta(x_k,y_k)+R_k-w_k\|^2]+\frac{m}{t^2}\eta_k^2\sigma_f(d_2)+\frac{2}{3\eta_k}(\|R_k\|^2+\|R_{k+1}\|^2)\\ &+\frac{2}{3}L_0^2\eta_k(\|x_k-\tilde{x}_{k+1}\|^2+\|y_k-\tilde{y}_{k+1}\|^2)\end{aligned} \tag{100}$$

where the last inequality holds by $\frac{9}{2}\gamma \leq c_1 \leq \frac{m^{1/2}}{t}$.

$$\begin{aligned}
&\mathbb{E}[\|\nabla g(x_{k+1}, y_{k+1}) - v_{k+1}\|^2] - \mathbb{E}[\|\nabla g(x_k, y_k) - w_k\|^2] \\
\leq& -\beta\mathbb{E}[\|\nabla g(x_k, y_k) - v_k\|^2] + \frac{2L_g^2}{\beta}\eta_k^2(\|x_k - \tilde{x}_{k+1}\|^2 + \|y_k - \tilde{y}_{k+1}\|^2) \\
\leq& -c_2\eta_k\mathbb{E}[\|\nabla g(x_k, y_k) - v_k\|^2] + \frac{2L_g^2}{c_2}\eta_k(\|\tilde{x}_k - x_{k+1}\|^2 + \|\tilde{y}_k - y_{k+1}\|^2) \\
\leq& -\frac{125L_0^2}{3\mu_g^2}\eta_k\mathbb{E}[\|\nabla g(x_k, y_k) - v_k\|^2] + \frac{6\mu_g^2 L_g^2}{125L_0^2}\eta_k(\|x_k - \tilde{x}_{k+1}\|^2 + \|y_k - \tilde{y}_{k+1}\|^2)
\end{aligned} \tag{101}$$

where the last inequality hold by $\frac{125L_0^2}{3\mu_g^2} \leq c_2 \leq \frac{m^{1/2}}{t}$.

In addition, we have

$$\begin{aligned}
&F_\delta(x_{k+1}) - F_\delta(x_k) \\
\leq& \eta_k\gamma c_l\left(2L_1^2(\frac{d_2}{\delta})\|y^*(x_k) - y_k\|^2 + 2\|\nabla f_\delta(x_k, y_k) - w_k\|^2\right) - \frac{\eta_k}{2\gamma c_l}\|\tilde{x}_{k+1} - x_k\|^2 \\
\leq& 2\eta_k\gamma c_l L_0^2\|y^*(x_k) - y_k\|^2 + 2\eta_k\gamma c_l\|\nabla f_\delta(x_k, y_k) - w_k\|^2 - \frac{\eta_k}{2\gamma c_l}\|\tilde{x}_{k+1} - x_k\|^2 \\
\leq& 2\eta_k\gamma c_l L_0^2\|y^*(x_k) - y_k\|^2 + 4\eta_k\gamma c_l\|\nabla f_\delta(x_k, y_k) - w_k - R_k\|^2 + 4\eta_k\gamma c_l\|R_k\|^2 - \frac{\eta_k}{2\gamma c_l}\|\tilde{x}_{k+1} - x_k\|^2
\end{aligned} \tag{102}$$

We can also have

$$\begin{aligned}
&\|y_{k+1} - y^*(x_{k+1})\|^2 - \|y^*(x_k) - y_k\|^2 \\
\leq& -\frac{\eta_k\tau\mu_g c_u}{4}\|y^*(x_k) - y_k\|^2 + \frac{25\eta_k\tau c_u}{6\mu_g}\|\nabla_y g(x_k, y_k) - v_k\|^2 \\
&- \frac{3\eta_k}{4}\|\tilde{y}_{k+1} - y_k\|^2 + \frac{25L_y^2\eta_k}{6\tau\mu_g c_u}\|x_k - \tilde{x}_{k+1}\|^2
\end{aligned} \tag{103}$$

Then, we define a Lyapunov function, for any $k \geq 1$,

$$\begin{aligned}
&\Phi_{k+1} \\
=& \mathbb{E}[F_\delta(x_{k+1}) + \frac{10L_0^a c_l}{\tau\mu_g c_u}\|y_{k+1} - y^*(x_{k+1})\|^2 + c_l(\|w_{k+1} - \bar{\nabla} f_\delta(x_{k+1}, y_{k+1}) - R_{k+1}\|^2 \\
&+ \|\nabla_y g(x_{k+1}, y_{k+1}) - v_{k+1}\|^2)]
\end{aligned} \tag{104}$$

We have

$$
\begin{aligned}
&\Phi_{k+1}-\Phi_k\\
=&\mathbb{E}[F_\delta(x_{k+1})-F_\delta(x_k)]+\frac{10L_0^a c_l}{\tau\mu c_u}\mathbb{E}[\|y_{k+1}-y^*(x_{k+1})\|^2-\|y_k-y^*(x_k)\|^2]\\
&+c_l\mathbb{E}[\|w_{k+1}-\bar{\nabla}f(x_{k+1},y_{k+1})-R_{k+1}\|^2-\|w_k-\bar{\nabla}f(x_k,y_k)-R_k\|^2]\\
&+c_l\mathbb{E}[\|\nabla_y g(x_{k+1},y_{k+1})-v_{k+1}\|^2-\|\nabla_y g(x_k,y_k)-v_k\|^2]\\
\le&2\eta_k\gamma c_l L_0^2\|y^*(x_k)-y_k\|^2+4\eta_k\gamma c_l\|\nabla f_\delta(x_k,y_k)-w_k-R_k\|^2+4\eta_k\gamma c_l\|R_k\|^2-\frac{\eta_k}{2\gamma c_l}\|\tilde{x}_{k+1}-x_k\|^2\\
&+\frac{10L_0^a c_l}{\tau\mu_g c_u}(-\frac{\eta_k\tau\mu_g c_u}{4}\|y^*(x_k)-y_k\|^2+\frac{25\eta_k\tau c_u}{6\mu_g}\|\nabla_y g(x_k,y_k)-v_k\|^2-\frac{3\eta_k}{4}\|\tilde{y}_{k+1}-y_k\|^2+\frac{25L_y^2\eta_k}{6\tau\mu_g c_u}\|x_k-\tilde{x}_{k+1}\|^2)\\
&+c_l(-\frac{9}{2}\gamma\eta_k\mathbb{E}[\|\nabla f_\delta(x_k,y_k)+R_k-w_k\|^2]+\frac{m}{t^2}\eta_k^2\sigma_f(d_2)+\frac{2}{3\eta_k}(\|R_k\|^2+\|R_{k+1}\|^2)\\
&+\frac{2}{3}L_0^2\eta_k(\|x_k-\tilde{x}_{k+1}\|^2+\|y_k-\tilde{y}_{k+1}\|^2))\\
&+c_l(-\frac{125L_0^2}{3\mu_g^2}\eta_k\mathbb{E}[\|\nabla g(x_k,y_k)-v_k\|^2]+\frac{6\mu_g^2L_g^2}{125L_0^2}\eta_k(\|x_k-\tilde{x}_{k+1}\|^2+\|y_k-\tilde{y}_{k+1}\|^2))\\
\le&(2\eta_k\gamma c_l L_0^2-\frac{5L_0^a c_l\eta_k}{2})\|y^*(x_k)-y_k\|^2+(4\eta_k\gamma c_l-\frac{9c_l\gamma}{2}\eta_k)\|\nabla f_\delta(x_k,y_k)-w_k-R_k\|^2\\
&+(\frac{125L_0^a c_l\eta_k}{3\mu_g^2}-c_l\frac{125L_0^2}{3\mu_g^2}\eta_k)\|\nabla_y g(x_k,y_k)-v_k\|^2\\
&+\left(-\frac{\eta_k}{2\gamma c_l}+\frac{125L_0^aL_y^2c_l\eta_k}{3\tau^2\mu_g^2c_u^2}+(\frac{2}{3}L_0^2+\frac{6\mu_g^2L_g^2}{125L_0^2})\eta_k c_l\right)\|\tilde{x}_{k+1}-x_k\|^2\\
&+\left(-\frac{15L_0^a c_l\eta_k}{2\tau\mu_g c_u}+(\frac{2}{3}L_0^2+\frac{6\mu_g^2L_g^2}{125L_0^2})\eta_k c_l\right)\|\tilde{y}_{k+1}-y_k\|^2\\
&+4\eta_k\gamma c_l\|R_k\|^2+\frac{2\gamma c_l}{3\eta_k}(\|R_k\|^2+\|R_{k+1}\|^2)+\frac{m}{t^2}\gamma c_l\eta_k^2\sigma_f(d_2)\\
\le&-\frac{L_0^2\gamma c_l\eta_k}{2}\|y^*(x_k)-y_k\|^2-\frac{\gamma c_l}{2}\eta_k\|\nabla f_\delta(x_k,y_k)-w_k-R_k\|^2\\
&-\frac{\eta_k}{4\gamma c_l}\|\tilde{x}_{k+1}-x_k\|^2+4\eta_k\gamma c_l\|R_k\|^2+\frac{2\gamma c_l}{3\eta_k}(\|R_k\|^2+\|R_{k+1}\|^2)\\
&+\frac{m}{t^2}\gamma c_l\eta_k^2\sigma_f(d_2)
\end{aligned}
\tag{105}
$$

where the last inequality is due to $0\le a\le 2$,$\gamma\le\min\{\frac{1}{L_0^{2-a}},\frac{1}{4c_l\left(\frac{125L_0^aL_y^2c_l}{3\tau^2\mu_g^2c_u^2}+(\frac{2}{3}L_0^2+\frac{6\mu_g^2L_g^2}{125L_0^2})c_l\right)},\frac{2m^{1/2}}{9t}\}$, $0<\tau\le\frac{15L_0^a}{2\mu_g c_u\left(\frac{2}{3}L_0^2+\frac{6\mu_g^2L_g^2}{125L_0^2}\right)}$ and $L_0>1$.

Then, rearranging the above inequality, we have

$$
\begin{aligned}
&\frac{\gamma c_l\eta_k}{4}\left(2L_0^2\|y^*(x_k)-y_k\|^2+2\|\nabla f_\delta(x_k,y_k)-w_k-R_k\|^2+\|R_k\|^2+\frac{1}{\gamma^2c_l^2}\|\tilde{x}_{k+1}-x_k\|^2\right)\\
\le&\frac{17}{4}\eta_k\gamma c_l\|R_k\|^2+\frac{2\gamma c_l}{3\eta_k}(\|R_k\|^2+\|R_{k+1}\|^2)+\frac{m}{t^2}\gamma c_l\eta_k^2\sigma_f(d_2)+\Phi_k-\Phi_{k+1}
\end{aligned}
\tag{106}
$$

Taking the average over $k=1,\cdots,K$ on both sides and using $\eta_k\ge\eta_K$, $Q=\frac{1}{\mu_g\eta}\ln\frac{C_{gxy}C_{fy}K}{\mu_g}$, $\eta_k=\frac{t}{(m+k)^{1/2}}$ and

$\Phi_1 = \mathbb{E}[F_\delta(x_1) + \frac{10L_0^a c_l}{\tau\mu_g c_u}\|y_1 - y^*(x_1)\|^2 + c_l(\|w_1 - \bar{\nabla} f_\delta(x_1, y_{k+1}) - R_1\|^2 + \|\nabla_y g(x_1, y_1) - v_1\|^2)]$, we have

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\frac{1}{4}\left(2L_0^2\|y^*(x_k) - y_k\|^2 + 2\|\nabla f_\delta(x_k, y_k) - w_k - R_k\|^2 + \|R_k\|^2 + \frac{1}{\gamma^2 c_l^2}\|\tilde{x}_{k+1} - x_k\|^2\right)] \\
\leq &\frac{1}{K\eta_K}\left(\frac{\Phi_1 - \Phi^*}{\gamma c_l} + \frac{17}{4K^2}\sum_{k=1}^{K}\eta_k + \frac{4}{3K^2}\sum_{k=1}^{K}\frac{1}{\eta_k} + (\frac{m}{t^2}\sigma_f(d_2))\sum_{k=1}^{K}\eta_k^2\right) \\
\leq &\frac{(m+K)^{1/2}}{Kt}\left(\frac{\Phi_1 - \Phi^*}{\gamma c_l} + \frac{17t}{4K^2}(m+K)^{1/2} + \frac{4}{3tK^2}(m+K)^{3/2} + (m\sigma_f(d_2))t^2\ln(m+K)\right)
\end{aligned} \tag{107}$$

According to the Jesen's inequality, we have

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\frac{1}{2}\left(\sqrt{2}L_0^2\|y^*(x_k) - y_k\| + \sqrt{2}\|\nabla f_\delta(x_k, y_k) - w_k - R_k\| + \|R_k\| + \frac{1}{\gamma c_l}\|\tilde{x}_{k+1} - x_k\|\right)] \\
\leq &\left(\frac{4}{K}\sum_{k=1}^{K}\frac{1}{4}\left(2L_0^2\|y^*(x_k) - y_k\|^2 + 2\|\nabla f_\delta(x_k, y_k) - w_k - R_k\|^2 + \|R_k\|^2 + \frac{1}{\gamma^2 c_l^2}\|\tilde{x}_{k+1} - x_k\|^2\right)\right)^{1/2} \\
\leq &\frac{2(m+K)^{1/4}}{\sqrt{Kt}}\sqrt{\frac{\Phi_1 - \Phi^*}{\gamma c_l} + \frac{17t}{4K^2}(m+K)^{1/2} + \frac{4}{3tK^2}(m+K)^{3/2} + (m\sigma_f(d_2))t^2\ln(m+K)} \\
\leq &\frac{2m^{1/4}\sqrt{G}}{\sqrt{Kt}} + \frac{2\sqrt{G}}{(Kt)^{1/4}}
\end{aligned} \tag{108}$$

where $G = \frac{\Phi_1 - \Phi^*}{\gamma c_l} + \frac{17t}{4K^2}(m+K)^{1/2} + \frac{4}{3tK^2}(m+K)^{3/2} + (m\sigma_f(d_2))t^2\ln(m+K)$.

Let $\mathcal{G}\left(x_k, \nabla F_\delta(x_k), \hat{\gamma}\right) = \frac{1}{\hat{\gamma}}\left(x_k - \mathcal{P}_{\mathcal{X}}\left(x_k - \hat{\gamma}\nabla F_\delta(x_k)\right)\right)$

$$\begin{aligned}
&\|\nabla F_\delta(x_k)\| \\
=&\|\mathcal{G}\left(x_k, \nabla F_\delta(x_k), \hat{\gamma}\right)\| \\
\leq&\|\mathcal{G}\left(x_k, \nabla F_\delta(x_k), \hat{\gamma}\right) - \mathcal{G}(x_t, w_k, \hat{\gamma})\| + \|\mathcal{G}(x_t, w_k, \hat{\gamma})\| \\
\leq&\|\nabla F_\delta(x_k) - w_k\| + \|\mathcal{G}(x_t, w_k, \hat{\gamma})\| \\
\leq&\|w_k - \nabla f(x_k, y_k) - R_k\| + \|R_k\|^2 + L_0\|y^*(x_k) - y_k\| + \frac{1}{\gamma c_l}\|x_k - \tilde{x}_{k+1}\|.
\end{aligned} \tag{109}$$

Finally, we can obtain

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\frac{1}{2}\|\nabla F_\delta(x_k)\|] \leq \frac{2m^{1/4}\sqrt{G}}{\sqrt{Kt}} + \frac{2\sqrt{G}}{(Kt)^{1/4}}. \tag{110}$$

Since the random count $r \in \{1, \cdot, K\}$ is uniformly sampled, we have

$$\mathbb{E}[\frac{1}{2}\|\nabla F_\delta(x_r)\|] = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\frac{1}{2}\|\nabla F_\delta(x_k)\|] \leq \frac{2m^{1/4}\sqrt{G}}{\sqrt{Kt}} + \frac{2\sqrt{G}}{(Kt)^{1/4}}. \tag{111}$$

By Proposition 3.9, we have $\nabla F_\delta(x_r) \in \bar{\partial}_\delta F(x_r)$. This implies that

$$\min\{\|h\| : h \in \bar{\partial}_\delta F(x_r)\} \leq \mathbb{E}[\|\nabla F_\delta(x_r)\|] \leq \frac{4m^{1/4}\sqrt{G}}{\sqrt{Kt}} + \frac{4\sqrt{G}}{(Kt)^{1/4}}. \tag{112}$$

That completes the proof.

□