Bilevel Optimization with Coupled Decision-Dependent Distributions

Songtao Lu [1]

Abstract

Bilevel optimization has gained significant popularity in recent years due to its ability to formulate various machine learning problems. For instance, in meta-learning, the upper-level (UL) problem offers a good initialization for the lower-level (LL) model to facilitate adaptation. However, the decision variables can impact data features and outcomes, leading to the phenomenon known as performativity. In this work, we investigate the inclusion of decision-dependent distributions in bilevel optimization. Specifically, we consider the scenarios where the UL data distribution depends on the LL optimization variable, and the LL data distribution also depends on the UL decision variable. We first establish sufficient conditions for the existence of performatively stable (PS) solutions in this class of bilevel problems. Also, we propose efficient stochastic algorithms to find the PS point with theoretical convergence rate analysis and discuss the theoretical optimality of the obtained solution. Our theoretical analysis is corroborated through a series of numerical experiments, wherein we evaluate the performance of the bilevel performative prediction algorithms alongside non-performative counterparts in the context of meta strategic learning problems.

1. Introduction

In this work, we consider the following class of bilevel performative prediction (PP) problems, where the data distribution at each level is dependent on the decision variable from the other level:

$$\mathbb{P}: \quad \min_{x} \quad \mathbb{E}_{Z\sim\mathcal{D}_x(y^*(x))} f(x, y^*(x); Z) \tag{1a}$$

$$\text{s.t.} \quad y^*(x) = \arg\min_{y} \mathbb{E}_{Z\sim\mathcal{D}_y(x)} \ell(x, y; Z) \tag{1b}$$

where x, y respectively denote the decision/optimization variables at the upper-level (UL) and lower-level (LL) of this problem, loss functions $f(x, y^*(x); Z)$ and $\ell(x, y; Z)$ are smooth, $y^*(x)$ represents the optimal solution of the LL problem given x, $\mathcal{D}_y(x)$ stands for the LL data distribution depended on the UL variable x, and similarly $\mathcal{D}_x(y^*(x))$ for the UL data distribution that is dependent on the LL loss function through the optimal LL decision variable $y^*(x)$.

Motivation of This Work. One of the most prominent machine learning models that can be formulated using bilevel optimization is model-agnostic meta-learning (MAML) (Finn et al., 2017; Rajeswaran et al., 2019). MAML has been successful in addressing the challenge of data distribution shifts between seen and unseen tasks, particularly in few-shot learning scenarios. In MAML, the meta-learner at the UL explores invariant features such as good initializations that can be utilized to handle unseen tasks, while the individual learners at the LL focus on fitting personalized data. However, in many real-world problems that fall under the umbrella of PP, such as election forecasts, financial markets, online advertising, and traffic predictions, the decision variables can heavily influence the data distribution (Perdomo et al., 2020). Hence, it becomes crucial to incorporate decision-dependent distributions into the framework of meta-learning, as both levels of the optimization problem are tightly intertwined through their mutual interaction. This motivates the inclusion of cross-level dependence between decision variables and data distributions in bilevel optimization.

Major Challenges in Bilevel PP. In contrast to single-level PP problems, solving bilevel optimization problems involves generating two sequences, each used for minimizing the loss function at its respective level. Due to the interdependence between decision variables and data distributions across levels, it is unclear whether changes in the data-generating distribution occur in a competitive or collaborative manner during the play of this performative Stackelberg game. Consequently, demonstrating the convergence rate of the bilevel stochastic algorithm is

[1] IBM Research, Thomas J. Watson Research Center, Yorktown Heights, New York 10598, USA. Correspondence to: Songtao Lu <songtao@ibm.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

highly challenging. It necessitates the construction of a Lyapunov function that accounts for the data distribution's dependence and satisfies a descent or contraction property. In addition, the theoretical analysis requires decoupling the stochastic error terms between the two levels.

Main Contributions of This Work. In this work, we provide the formal definitions of performatively optimal (PO) and performatively stable (PS) points in bilevel optimization and establish the conditions for the existence of the bilevel performatively stable (BPS) points using the bilevel repeated risk minimization (Bi-RRM) method. Building upon the existence of the BPS points, we propose an efficient bilevel stochastic gradient descent (Bi-SGD) method and derive sufficient conditions for Bi-SGD to obtain the BPS point. Our theoretical analysis reveals that Bi-SGD achieves comparable iteration and sample complexities to traditional (bilevel) stochastic gradient descent (SGD) for finding optimal solutions in non-performative strongly convex (bilevel) problems. To validate our theoretical results and demonstrate the stability of Bi-SGD under data distribution shifts, we conduct experiments on both synthetic and real data sets. The results underscore the efficacy of Bi-RRM and Bi-SGD. The main contributions of this work are highlighted as follows.

- **Bilevel performative prediction**. To the best of our knowledge, our proposed bilevel PP model is the first to incorporate decision-dependent data distributions into the learning process. This unique approach distinguishes our work from existing models in the field.
- **Theoretical analysis**. Our theoretical results provide novel insights into the existence of BPS points in the bilevel PP model. We demonstrate, for the first time, that when the sensitivity parameters of the data distribution at both levels fall below a certain threshold, a unique BPS point exists. Additionally, we prove that the bilevel SGD algorithm is capable of achieving the PS point at a rate of $\mathcal{O}(1/T)$, which is on par with the standard non-performative learning setting. Here, T represents the total number of iterations.
- **Applications to meta strategic learning**. We present the results of applying the bilevel PP model to meta strategic learning and showcase the importance of incorporating the decision-dependent distribution in the algorithm design in terms of improving testing accuracy.

2. Background and Related Work

Performative Prediction (PP). There is a significant body of research focusing on various aspects of PP, including model and algorithm design, convergence analysis, and quantification of generalization performance. The concept of PP was first introduced in (Perdomo et al., 2020), which addresses the strategic feedback effect in single-level risk minimization problems. The work also proposes measures to capture performative optimality and stability in this setting, where the data distribution is induced by the predictive model. It is shown that under certain conditions on the loss function (such as smoothness and strong convexity), the PS point can be achieved by utilizing repeated risk minimization (RRM) or gradient descent (GD) on the performative risk when the changes in the data distribution are not significant or controllable in terms of the Wasserstein-1 distance.

Based on this work, more efficient SGD methods have been proposed in (Mendler-Dünner et al., 2020) to find the PS points by considering the trade-off between the frequency of model deployment and the stochastic update of the model. In (Drusvyatskiy & Xiao, 2022), several families of classical stochastic optimization algorithms, such as clipped gradient and proximal point, have been studied for PP with rigorous convergence rate analysis. These works highlight the main challenge in proving algorithm convergence, which lies in quantifying the bias introduced by distribution shifts after model deployment. Also, it is worth noting that the convergence results in these seminal works rely on the strong convexity assumption of the loss functions. Recent works have attempted to relax this assumption to the convex case (Miller et al., 2021) or the weakly convex case (Zhao, 2022), but they require additional assumptions, such as the mixture dominance condition or other notions of Lipschitz distributions (Mofakhami et al., 2022), to ensure the well-behaved nature of the distribution map.

Given the dynamic nature of online data sequences, PP naturally extends to model richer classes of online learning problems, enabling the advancement of existing algorithms to adapt to changing environments and facilitating the design of new algorithms to address distribution shifts. For instance, online projected gradient descent has been applied to optimize the charging of a fleet of electric vehicles, where the time-varying costs depend on random variables with decision-dependent distributions (Wood et al., 2021). The proximal stochastic gradient method has been employed to tackle online tracking problems, where the dynamics are jointly dependent on both time and decision variables, and its non-asymptotic convergence behavior under the time drift is analyzed in (Cutler et al., 2021). Furthermore, performative feedback has been utilized in algorithm design to construct confidence bounds on the risk of unexplored models and guide exploration (Jagadeesan et al., 2022). If the exact gradient of the temporal drift function over time is accessible, the underlying changes in dynamics can be approximated in an online fashion. Based on this observation, the predictor-corrector method is proposed for time-varying stochastic optimization (Maity et al., 2022).

Table 1. Comparison of the existing representative works that are related to PP and bilevel optimization, where Opt.: optimization, Cond.: the existence condition of the PS points, Gap: the maximum distance between the PS and PO points, Deploy: the single-loop/double-loop structure of deploying the algorithm/method.

Algorithms	Opt. Framework	Cond.	Gap	Deploy	Rate
RRM and GD (Perdomo et al., 2020)	single-level	$\mathcal{O}(\varepsilon)$	$\mathcal{O}(\varepsilon)$	single	$\mathcal{O}(\log(1/T))$
SGD (Mendler-Dünner et al., 2020)	single-level	$\mathcal{O}(\varepsilon)$	$\mathcal{O}(\varepsilon)$	single	$\mathcal{O}(1/T)$
Multi-PfD (Li et al., 2022b)	single-level+consensus	$\mathcal{O}(\varepsilon_{\text{avg}})$	n/a	single	$\mathcal{O}(1/T)$
BSA (Ghadimi & Wang, 2018)	bilevel	n/a	n/a	double	$\mathcal{O}(1/T)$
STSA (Shen & Chen, 2022)	bilevel	n/a	n/a	single	$\mathcal{O}(1/T)$
Bi-RRM (This work)	bilevel	$\mathcal{O}(\varepsilon_x\varepsilon_y + \varepsilon_y)$	$\mathcal{O}(\varepsilon_x(1+\varepsilon_y))$	double	$\mathcal{O}(\log(1/T))$
Bi-SGD (This work)	bilevel	$\mathcal{O}(\varepsilon_x+\varepsilon_x\varepsilon_y+\varepsilon_y)$	$\mathcal{O}(\varepsilon_x(1+\varepsilon_y))$	single	$\mathcal{O}(1/T)$

In addition to the applications of PP in online learning, there is another line of research that focuses on PP for performative reinforcement learning, where the policy affects both the underlying reward and the transition kernel of the Markov chain (Mandal et al., 2022). Subsequently, the convergence analysis of state-dependent PP algorithms has been investigated under the Markov transition model (Li & Wai, 2022; Roy et al., 2022; Brown et al., 2022) as well as general stateful performative dynamics (Izzo et al., 2022). When prior knowledge about the decision-dependent distribution is available, more informative PP algorithms can be designed using performative gradient descent, which can improve the performance of classic gradient-based methods under distribution shifts (Izzo et al., 2021). Apart from the optimization perspective, recent research has also focused on identifying the causal effect of predictions (Mendler-Dünner et al., 2022), exploring the performative power by measuring the relationship between the decision maker and the population (Hardt et al., 2022), and analyzing outcome performativity with an emphasis on the performative effects of decisions on the conditional distribution rather than the traditional overall distribution (Kim & Perdomo, 2022).

Two (multiple) Players Game. The decision-making process and strategic response in PP can be also viewed as a two-player game, where the data reacts based on the decision variables to maximize a utility function. PP assumes that this interaction occurs simultaneously. In (Zrnic et al., 2021), the order of play in strategic classification between the decision maker and strategic agents is further examined, revealing that the update frequency determines the roles of the leader and follower. Furthermore, algorithm design for PP in zero-sum games has been developed, along with corresponding convergence analysis (Wood & DallAnese, 2022; Maheshwari et al., 2022). In a multi-agent setting, decision makers can work in either a competitive or collaborative manner. In (Narang et al., 2022), a competitive multi-player performative prediction setting is introduced, where each local player aims to minimize their performative risk under the joint action space, considering that the local data distribution is influenced by the decisions of all players. It is shown that using SGD at each player can effectively find the Nash equilibrium of this problem. In (Li et al., 2022b), a consensus-based multi-agent performative prediction (Multi-PfD) is considered, where all agents connected through a communication network seek a global PS solution by interacting with local strategic data. Additionally, it is demonstrated that the existence condition regarding the ε-sensitivity parameter can be relaxed from ε (in the single-agent setting) to ε_{avg}, where ε_{avg} represents the average sensitivity parameter across the entire network.

Bilevel Optimization and Meta-Learning. Model transferability is a topic of significant interest in decision-dependent learning systems (Liu et al., 2021b). MAML has emerged as a powerful tool for enhancing the generalization performance of trained models when faced with new tasks in both supervised learning and reinforcement learning settings (Liu et al., 2019; Rajeswaran et al., 2019). The convergence rate of MAML and regret analysis have been characterized in previous studies (Balcan et al., 2019; Fallah et al., 2020), along with investigations into generalization errors (Chua et al., 2021; Chen et al., 2022). As mentioned earlier, MAML can be seen as a special case of bilevel optimization (Franceschi et al., 2018) or a Stackelberg game (Fiez et al., 2020). Also, bilevel optimization can be applied to various other machine learning problems, including multi-task AUC maximization (Hu et al., 2022), meta causal discovery (Lu & Gao, 2023), data hyper-cleaning (Franceschi et al., 2018), etc.

Consequently, solving the bilevel optimization problem becomes crucial. One of the earliest approaches for deterministic convex bilevel problems is the bilevel gradient sequential averaging method (Sabach & Shtern, 2017), while for stochastic bilevel problems, the bilevel stochastic approximation (BSA) method is commonly adopted (Ghadimi & Wang, 2018). BSA solves the LL problem by iteratively updating its variables in an inner loop and then switches to optimize the UL variables. The theoretical convergence rates of BSA have been provided in (Ghadimi & Wang, 2018) for strongly convex, convex, and nonconvex UL loss functions when the LL loss function is strongly convex. Building upon this, subsequent research has focused on improving the iteration or sample complexity of BSA. Examples include the development of single-time scale or multi-sequence single-timescale (STSA) bilevel algorithms (Chen et al., 2021; Shen & Chen, 2022).

When the LL objective function is strongly convex, it is possible to derive a closed form of the UL gradient. However, computing this hyper-gradient requires the inversion of the LL Hessian matrix. To enhance computation efficiency, multiple efficient bilevel algorithms and techniques have been proposed, including approximate implicit differentiation (Ji et al., 2021), adaptive stochastic algorithms (Huang & Huang, 2021), and an adaptation of the well-known SAGA algorithm (Dagréou et al., 2022; Li et al., 2022a). Besides, generalized bilevel optimization solvers have been developed by relaxing the strong convexity assumption of the LL loss function (Ye et al., 2022; Liu et al., 2022). A comprehensive survey paper (Liu et al., 2021a) provides an overview of these research areas, and Table 1 presents a summary of the representative works related to both PP and bilevel optimization.

3. Bilevel Performative Optimality

In this work, we consider the scenario where the data distribution over the features and outcomes at each level of the problem depends on the decision variables at the other level. Consequently, the performance evaluation of this type of two-player performative model is based on the expected loss over the distributions induced at both levels. To be more precise, in contrast to existing single-level PP models (Perdomo et al., 2020), our study focuses on a bilevel model that aims to find the following optimal points through a performative Stackelberg game.

Definition 1. *(Bilevel performative optimality and risk). A point x_{O} is bilevel performatively optimal (BPO) if it satisfies*

$$x_{\mathrm{O}} = \arg\min_x F(x) \tag{2}$$

where $F(x) \triangleq \underset{Z\sim\mathcal{D}(y^*(x))}{\mathbb{E}}[f(x, y^*(x); Z)]$ *is defined as the UL performative risk,* $y^*(x)$ *is defined in* (1b)*, and* $\mathbb{E}_{Z\sim\mathcal{D}_y(x)}\ell(x, y; Z)$ *is defined as the LL performative risk.*

PS is another well-established notion of PP, which refers to a point that attains the global optimal solution of the optimization problem considering the data distributions induced by that point itself. A formal definition of the bilevel PS point can be expressed as follows.

Definition 2. *(Bilevel performative stability and decoupled risk). A point x_{S} is bilevel performatively stable (BPS) if it satisfies*

$$x_{\mathrm{S}} \triangleq \arg\min_x \underset{Z\sim\mathcal{D}(y^*(x,x_{\mathrm{S}}))}{\mathbb{E}}[f(x, y^*(x, x_{\mathrm{S}}); Z)] \tag{3a}$$

$$\text{s.t. } y^*(x, x_{\mathrm{S}}) = \arg\min_y \underset{Z\sim\mathcal{D}(x_{\mathrm{S}})}{\mathbb{E}}[\ell(x, y; Z)]. \tag{3b}$$

Also, let $DR(x, x') \triangleq \mathbb{E}_{Z\sim\mathcal{D}(y^*(x,x'))}[f(x, y^*(x, x'); Z)]$ *be the decoupled bilevel performative risk.*

It is clear that $x_{\mathrm{S}} = \arg\min_x \mathrm{DR}(x, x_{\mathrm{S}})$ and $x_{\mathrm{O}} = \arg\min_x \mathrm{DR}(x, x)$. The concept of performative stability revolves around the idea of a fixed point in risk minimization, wherein the learned model minimizes the risk on the data distribution that arises from its own deployment (Perdomo et al., 2020). This property validates the optimality of the closed-loop training strategy and serves as motivation for incorporating PP in the meta-learning setting. In this case, the objective of both levels of learning is to identify the permutation invariant space after the model has been deployed.

It is evident that the decision-dependent distribution plays a crucial role in bridging the gap between model deployment and model parameter optimization. Similar to the concept of Lipschitz continuity used in the field of optimization to quantify function changes, the literature (Perdomo et al., 2020) has introduced the notion of ε-sensitivity. This measure is employed to assess the variations in decision-dependent distributions caused by changes in the decision variables.

Definition 3. *(ε-sensitive) A distribution map $\mathcal{D}(\cdot)$ is ε-sensitive if all x, x':*

$$\mathcal{W}_1(\mathcal{D}(x), \mathcal{D}(x')) \leq \varepsilon\|x - x'\|_2 \tag{4}$$

where $\mathcal{W}_1$ denotes the Wasserstein-1 distance (aka the earth mover's distance) between two distributions.

Given the definitions of BPO, BPS, and ε-sensitivity, it is still not clear whether the BPO or BPS exists or not. Note that the decision variables compete to minimize their objective functions through the UL and LL optimization processes, along with the decision-dependent distributions. This coupling can lead to oscillations in the iterates generated by the optimization algorithms, making the convergence analysis challenging, as mentioned earlier. To address this, we first introduce Bi-RRM, which demonstrates the existence of the BPS point and provides insights into the iteration complexity of finding this point.

4. Existence of BPS Point

In this section, we will propose the Bi-RRM method for solving this bilevel PP problem (1).

4.1. Bilevel Repeated Risk Minimization (Bi-RRM)

The main idea of RRM involves a retraining procedure outlined as follows. It begins by solving the bilevel optimization problem using the data distribution induced by the decision variable x_r. Next, it updates the state to x_{r+1} using the obtained solution, and this process is repeated iteratively, with r denoting the index of the iterations. Mathematically, Bi-RRM performs the following recursive update:

$$
\begin{aligned}
x_{r+1} &= R(x_r) \\
&\triangleq \arg\min_{\varphi} \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(\varphi,x_r))} [f(\varphi, y^*(\varphi, x_r); Z)], \quad (5a) \\
\text{s.t. } y^*(\varphi, x_r) &= \arg\min_{y} \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x_r)} [\ell(\varphi, y; Z)] \quad (5b)
\end{aligned}
$$

where $R(\cdot)$ represents the one-step update of the RRM algorithm. The Bi-RRM assumes the existence of an oracle algorithm that can obtain the global optimal solution of the bilevel optimization problem with respect to the variable φ. This optimization problem takes into account both the UL and LL data distributions induced by the decision variable x at iteration r.

4.2. Theoretical Assumptions

Before showing the theoretical convergence result of Bi-RRM, we make the following blanket assumption for problem (1).

Without loss of generality, we assume that decision-dependent distributions at both UL and LL are ε-sensitive but with different constants.

Assumption 1. *Assume that the distribution maps are ε_x- and ε_y-sensitive at each level, namely, $\mathcal{W}_1(\mathcal{D}_x(x), \mathcal{D}_x(x')) \le \varepsilon_x\|x - x'\|_2$ and $\mathcal{W}_1(\mathcal{D}_y(y), \mathcal{D}_y(y')) \le \varepsilon_y\|y - y'\|_2$.*

Next, we assume the strong convexity and smoothness of the loss functions as follows.

Assumption 2. *(Strong convexity) Assume that loss functions $F(x)$ and $\ell(x, y)$ are respectively γ_x- and γ_y-strongly convex, namely,*

$$F(x) \ge F(x') + \nabla_x F(x')^T(x - x') + \frac{\gamma_x}{2}\|x - x'\|^2,$$

$$\ell(x, y) \ge \ell(x, y') + \nabla_y \ell(x, y')^T(y - y') + \frac{\gamma_y}{2}\|y - y'\|^2.$$

Assumption 3. *(Smoothness of the UL loss function and distribution) Assume that the loss function $f(x, y; Z)$ w.r.t. x, y is smooth and the gradients of $f(x, y; Z)$ w.r.t. Z are jointly L_f^z-Lipschitz continuous $\forall x, y$, namely,*

$$\|\nabla_x f(x, y; Z) - \nabla_x f(x, y; Z')\| \le L_f^z\|Z - Z'\|, \quad (6a)$$

$$\|\nabla_y f(x, y; Z) - \nabla_y f(x, y; Z')\| \le L_f^z\|Z - Z'\|. \quad (6b)$$

(Smoothness of the LL loss function and distribution) Similarly, we assume that loss function $\ell(x, y; Z)$ is smooth and the gradient of $\ell(x, y; Z)$ is L_ℓ^z-Lipschitz continuous $\forall x, y$, namely,

$$\|\nabla_y \ell(x, y; Z) - \nabla_y \ell(x, y; Z')\| \le L_\ell^z\|Z - Z'\|. \quad (7)$$

(Second-order smoothness of the LL loss function and distribution) Assume that loss function $\ell(x, y)$ is continuously twice differentiable and its the Jacobian and Hessian matrices are respectively $L_{\ell xy}^z$- and $L_{\ell yy}^z$-Lipschitz continuous $\forall x, y$, namely,

$$\|\nabla_{xy}^2 \ell(x, y; Z) - \nabla_{xy}^2 \ell(x, y; Z')\| \le L_{\ell xy}^z\|Z - Z'\|, \quad (8a)$$

$$\|\nabla_{yy}^2 \ell(x, y; Z) - \nabla_{yy}^2 \ell(x, y; Z')\| \le L_{\ell yy}^z\|Z - Z'\|. \quad (8b)$$

Assumption 4. *(Boundedness of the gradient and Jacobian) We assume that $\|\nabla_y f(x, y)\| \le C_f^y$ and $\|\nabla_{xy}^2 \ell(x, y)\| \le C_{\ell xy}$ $\forall x, y$.*

The assumptions of the Lipschitz continuity on data distributions are commonly used in the existing theoretical works of quantifying the convergence of the single-level PP algorithms (Mendler-Dünner et al., 2020; Drusvyatskiy & Xiao, 2022).

Remark 1. Under the strong convexity assumption of the LL objective function, we have the closed form of computing the UL gradient through the chain rule, which is $\nabla_x F(x) = \nabla_x f(x, y^*(x)) - \nabla_{xy}^2 \ell(x, y^*(x)) \left[\nabla_{yy}^2 \ell(x, y^*(x))\right]^{-1} \nabla_y f(x, y^*(x))$ (Ghadimi & Wang, 2018; Shen & Chen, 2022). It can be seen that the UL gradient involves the second-order derivatives of the LL loss function. Therefore, we assume the second-order continuity and boundedness of the loss function, which will be used for measuring the distribution changes with respect to the UL and LL variables. In the following, we will present the theoretical results and relegate the detailed proofs in the appendix.

4.3. Convergence of Bi-RRM

Theorem 1. *Suppose that A.1–A.4 hold and iterates $\{x_r, \forall r \ge 1\}$ are generated by the Bi-RRM method. Then,*

$$\|R(x) - R(x')\| \le (C_{xy}\varepsilon_x\varepsilon_y + C_y\varepsilon_y)\|x - x'\|, \forall x, x' \quad (9)$$

where

$$C_{xy} \triangleq \frac{C_{\ell xy}}{\gamma_x\gamma_y}\left(L_f^z + \frac{C_{\ell xy}L_f^z}{\gamma_y}\right), \quad (10a)$$

$$
\begin{aligned}
C_y \triangleq \frac{1}{\gamma_x\gamma_y}\Bigg(&C_f^y\left(L_{\ell xy}^z + \frac{L_{\ell yy}^z C_{\ell xy}}{\gamma_y}\right) \\
&+ L_\ell^z\left(\bar{L}_f^x + \frac{C_f^y L_{\ell xy}^y}{\gamma_y} + \frac{C_{\ell xy}}{\gamma_y}\left(L_f^y + \frac{C_f^y L_{\ell yy}^y}{\gamma_y}\right)\right)\Bigg). \quad (10b)
\end{aligned}
$$

If ε_x and ε_y satisfy

$$C_{xy}\varepsilon_x\varepsilon_y + C_y\varepsilon_y \le 1, \quad (11)$$

then the sequence generated by Bi-RRM converges to a unique BPS point x_{S} at a linear rate, namely,

$$\|x_r - x_{\mathrm{S}}\|_2 \le (C_{xy}\varepsilon_x\varepsilon_y + C_y\varepsilon_y)^r\|x_1 - x_{\mathrm{S}}\|.$$

This result claims that under A.1–A.4 when (11) is satisfied, there exists a global optimal BPS solution of problem (1).

Remark 2. If we set $\varepsilon_x = \varepsilon_y \triangleq \varepsilon$, it becomes evident that the condition (11) ensuring the linear convergence rate of Bi-RRM to the BPS solution is $\varepsilon < (2C_{xy})^{-1}(\sqrt{C_y^2 + 4C_{xy}} - C_y)$.

Remark 3. Another important observation is that ε_x and ε_y are coupled in a bilinear manner in the existence condition (11) of the BPS point. This means that if both values are large, the coupling term will be amplified, making it impossible for $C_{xy}\varepsilon_x\varepsilon_y + C_y\varepsilon_y$ to be less than 1. Additionally, it is worth noting that the asymmetry between C_{xy} and C_y is dependent on $C_{\ell xy}, L^y_{\ell xy}, L^z_{\ell xy}$. This observation aligns with intuition, as the Jacobian matrix reflects the coupling relationship between x and y.

5. Bilevel Performative Prediction Algorithms

However, the Bi-RRM method is computationally inefficient as it requires solving a bilevel optimization problem completely at each iteration, including assessing the full gradient. In order to address this issue, we propose a simple gradient-based stochastic bilevel algorithm to find the BPS point.

5.1. Bilevel Stochastic Gradient Descent (Bi-SGD)

To simplify our notation, we define $\ell(x, y) = \mathbb{E}_{Z\sim\mathcal{D}(x)}[\ell(x, y; Z)]$ and $f(x, y^*(x)) = \mathbb{E}_{Z\sim\mathcal{D}(y^*(x))}[f(x, y^*(x); Z)]$. Furthermore, we use $\ell(x_r, y_r; Z_y)$ to denote $\ell(x_r, y_r; Z)$, where $Z \sim \mathcal{D}_y(x_r)$, and $f(x_r, y_r; Z_x)$ to denote $f(x_r, y_r; Z)$, where $Z \sim \mathcal{D}_x(y_r)$. Then, the proposed bilevel SGD algorithm for minimizing both the UL and LL performative risks can be written concisely as follows.

$$y_{r+1} = y_r - \beta_r\widehat{\nabla}_y\ell(x_r, y_r; Z_y), \tag{12a}$$

$$x_{r+1} = x_r - \alpha_r\widehat{\nabla}_x f(x_r, y_r; Z_x), \tag{12b}$$

where α_r, β_r respectively denote the step sizes of the UL and LL learning processes, and $\widehat{\nabla}_y\ell(x_r, y_r; Z_y), \widehat{\nabla}_x f(x_r, y_r; Z_x)$ respectively represent the gradient estimates of $\nabla_y\ell(x_r, y_r)$ and $\overline{\nabla}_x f(x_r, y_r)$ with only utilizing a minibatch of *i.i.d.* data samples. Here, $\overline{\nabla}_x f(x, y)$ denotes the surrogate of $\nabla_x F(x)$, which simply replaces $y^*(x)$ in $\nabla_x F(x)$ by y.

Remark 4. If the full gradients of both the UL and LL loss functions can be obtained at each iteration, the Bi-SGD algorithm reduces to bilevel gradient descent (Bi-GD).

5.2. Theoretical Assumptions

To quantify the descent achieved by Bi-SGD after each round of updates, we need the following assumptions.

Assumption 5. *Assume that the gradient of loss function $f(x, y)$ w.r.t. x is Lipschitz continuous with constant L^x_f for x and $\bar{L}^x_f$ for y. Similarly, we assume that the gradient of $f(x, y)$ w.r.t. y is Lipschitz continuous with constant L^y_f for y and $\bar{L}^y_f$ for x, and loss function $\ell(x, y)$ is L_ℓ-smooth.*

Also, we assume that the Jacobian and Hessian matrices of loss function $\ell(x, y)$ are Lipschitz continuous with constants $L^x_{\ell xy}$ and $L^x_{\ell yy}$ for x and constants $L^y_{\ell xy}$ and $L^y_{\ell yy}$ for y.

Let $\mathcal{F}_r = \sigma\{y_1, x_1, \dots, y_r, x_r\}$ denote the filtration of the random variables up to iteration r, where $\sigma\{\cdot\}$ is the σ-algebra generated by the random variables.

Assumption 6. *(Quality of both the LL and UL gradient estimates) Assume that the LL gradient estimate is unbiased and with bounded variance, namely,*

$$\mathbb{E}[\widehat{\nabla}_y\ell(x_r, y_r; Z)|\mathcal{F}_r] = \nabla_y\ell(x_r, y_r), \tag{13a}$$

$$\mathbb{E}[\|\widehat{\nabla}_y\ell(x_r, y_r; Z) - \nabla_y\ell(x_r, y_r)\|^2|\mathcal{F}_r] \le \sigma^2_\ell \tag{13b}$$

Assume that the UL gradient estimate is biased and with bounded variance, namely,

$$\mathbb{E}[\widehat{\nabla}_x f(x_r, y_r; Z)|\mathcal{F}_r] \triangleq \overline{\nabla}_x f(x_r, y_r) + b_r, \tag{14}$$

$$\mathbb{E}[\|\widehat{\nabla}_x f(x_r, y_r; Z) - \overline{\nabla}_x f(x_r, y_r) - b_r\|^2|\mathcal{F}_r] \le \sigma^2_f,$$

where b_r denotes the bias term and $\|b_r\| \le \delta_r, \forall r$.

Remark 5. The bias term b_r primarily arises from the estimation of the inverse Hessian matrix during the stochastic approximation of the UL gradient. In (Ghadimi & Wang, 2018), it has been demonstrated that by utilizing the Hessian inverse approximation sampling method, $\widehat{\nabla}_x f(x_r, y_r; Z_x)$ can be accurately obtained. Additionally, it has been shown that the size of the resulting bias term exponentially decreases as the number of samples increases.

All the aforementioned assumptions regarding Lipschitz continuity and the quality of gradient estimates are standard in the convergence analysis for stochastic bilevel algorithms (Ghadimi & Wang, 2018; Dagréou et al., 2022; Chen et al., 2021; Shen & Chen, 2022). In the following section, we will present the theoretical results and provide detailed proofs in the appendix.

5.3. Convergence Rates of Bi-SGD and Bi-GD

Based on these mild assumptions, we can show the main theorem regarding the convergence rate of Bi-SGD as follows.

Theorem 2. *(Convergence Rate of Bi-SGD) Suppose that A.1-A.6 hold and iterates $\{x_r, y_r, \forall r \ge 1\}$ are generated by Bi-SGD. When the step sizes are chosen as $\alpha_r = \Theta(1/r)$, $\beta_r = \Theta(1/r)$ and ε_x and ε_y satisfy*

$$C_x\varepsilon_x + C_{xy}\varepsilon_x\varepsilon_y + C_y\varepsilon_y \le \frac{L_F^{\varepsilon_x,\varepsilon_y}}{4(L_F^{\varepsilon_x,\varepsilon_y} + \gamma_x)} \tag{15}$$

where

$$C_x \triangleq \left(L^z_f + \frac{C_{\ell xy}L^z_f}{\gamma_y}\right)\frac{C_{\ell xy}}{\gamma_x\gamma_y}, \tag{16}$$

then, it holds for any r that

$$\mathbb{E}\|x_r - x_{\mathrm{S}}\|^2 + \mathbb{E}\|y_r - y^*(x_{\mathrm{S}})\|^2 = \mathcal{O}\left(\frac{1}{r}\right) \tag{17}$$

where $L_F^{\varepsilon_x,\varepsilon_y}$ is linear in terms of ε_x and ε_y (the detailed expression of $L_F^{\varepsilon_x,\varepsilon_y}$ is shown in Lemma 3 in the appendix.) Moreover, we have

$$\lim_{r\to\infty}\|x_r - x_{\mathrm{S}}\|^2 \to 0, \quad \lim_{r\to\infty}\|y_r - y^*(x_{\mathrm{S}})\|^2 \to 0 \tag{18}$$

almost surely.

Remark 6. By comparing the conditions (11) and (15), it is evident that Bi-SGD imposes more stringent requirements, necessitating smaller values of ε_x and ε_y in order to find the BPS point.

Corollary 1. *(Convergence Rate of Bi-GD) Suppose that A.1-A.5 hold and iterates $\{x_r, y_r, \forall r \geq 1\}$ are generated by Bi-GD. When the step sizes are chosen as $\alpha_r = \Theta(1)$, $\beta_r = \Theta(1)$ and ε_x and ε_y satisfy* (15)*, it holds for any r that*

$$\mathbb{V}_{r+1} \leq \left(1 - \min\left(\frac{L_F^{\varepsilon_x,\varepsilon_y}\gamma_x\alpha_r}{2(L_F^{\varepsilon_x,\varepsilon_y}+\gamma_x)}, \frac{\gamma_y\beta_r}{2}\right)\right)\mathbb{V}_r \quad (19)$$

where $\mathbb{V}_r \triangleq \|x_r - x_{\mathrm{S}}\|^2 + \|y_r - y^*(x_{\mathrm{S}})\|^2$.

Remark 7. Corollary 1 claims that Bi-GD can achieve a linear convergence rate in converging to the BPS solution of problem (1), which is the same as Bi-RRM.

6. Relation between BPS and BPO Points

The aforementioned results demonstrate that Bi-SGD or Bi-GD (or Bi-RRM) can achieve the BPS point at a linear or sublinear rate. However, directly minimizing the performative risk to find the BPO points is a challenging task, even in the single-level case (Perdomo et al., 2020). Instead, we can estimate the distance between x_{O} and x_{S} by assuming that the loss function satisfies Lipschitz continuity with respect to the data distribution, namely, $|f(x, y^*(x); Z) - f(x, y^*(x); Z')| \leq L_z|Z - Z'|$, where L_z is a constant.

Theorem 3. *Suppose that A.1-A.4 hold and function $f(x, y^*(x); Z)$ is L_z-Lipschitz in Z. Then, for every BPS point and BPO point, the following relation holds.*

$$\|x_{\mathrm{O}} - x_{\mathrm{S}}\| \leq \frac{2\varepsilon_x L_z}{\gamma_x\gamma_y}\left(C_{\ell xy} + L_\ell^z\varepsilon_y\right). \quad (20)$$

Remark 8. Theorem 3 shows that the distance between x_{O} and x_{S} can be bounded by the values of ε_x and ε_y. This bound indicates that when ε_x and ε_y are small, x_{S} is very close to x_{O}. Furthermore, it suggests that x_{S} can serve as an approximation of x_{O}, with ε_x playing a more dominant role in the error bound compared to ε_y.

7. Numerical Experiments

In this section, we conduct experiments to evaluate the performance of Bi-RRM and Bi-SGD on the bilevel strategic classification problem and the meta strategic learning problem using both the synthetic and real data sets. In all the experiments, we adopt a linear utility function to model the reactions of the decision variables to the data, following the approach in (Perdomo et al., 2020). Specifically, for Bi-RRM, the feature vectors are shifted by $\varepsilon_x y^*(x_r)$ for Bi-RRM at UL ($\varepsilon_x y_r$ for Bi-SGD) and $\varepsilon_y x_r$ at LL. Additionally, we compare the convergence behavior of these algorithms under different choices of ε_x and ε_y.

7.1. Toy Example

Firstly, we consider a simple strategic bilevel problem as follows,

$$\min_{x\in\mathbb{R}^d} \mathop{\mathbb{E}}_{\{(a_i,b_i)\}\sim\mathcal{D}_x(y^*(x))} \|b_i - a_i^T x\| + \frac{\lambda_x}{2}\|x - y^*(x)\|^2$$

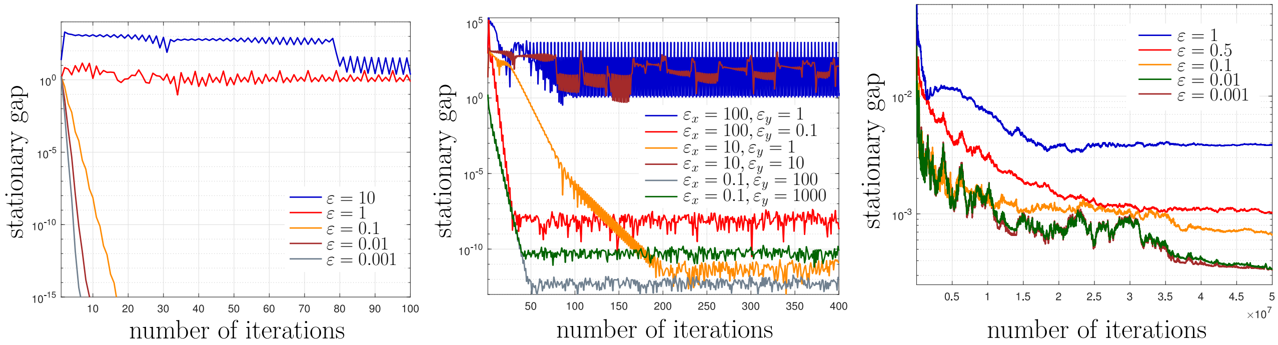
$$\text{s.t. } y^*(x) = \arg\min_{y\in\mathbb{R}^d} \mathop{\mathbb{E}}_{\{(c_i,d_i)\}\sim\mathcal{D}_y(x)} \|d_i - c_i^T y\| + \frac{\lambda_y}{2}\|y - x\|^2,$$

where both UL and LL objective functions are regression mismatch loss plus a quadratic regularization term. Here, b_i and d_i are generated respectively through the linear regression models, i.e., $b_i = a_i^T x^\dagger + n_x$ and $d_i = c_i^T y^\dagger + n_y, \forall i$, where $x^\dagger, y^\dagger$ and noise terms n_x, n_y are *i.i.d.* Gaussian random variables with zero mean and unit variance. The features a_i and c_i are *i.i.d.* Gaussian random variables with zero mean and variance 1 and 2 respectively. The regularization terms penalize the dissimilarity between the UL and LL variables, which appears commonly in the federated learning (T Dinh et al., 2020) and meta-learning (Rajeswaran et al., 2019) settings.

In our numerical experiments, we set the problem dimension as 5 and the total number of data samples as 50. The parameter ε represents both ε_x and ε_y. We choose $\lambda_x = \lambda_y = 1 \times 10^{-3}$, the minibatch size as 5, and use the same step size $1/\sqrt{r}$ for both α_r and β_r in Bi-SGD. In this example, we measure the optimality of the solutions based on the size of the UL gradient. Note that this problem is strongly convex and unconstrained, so a zero stationary gap indicates the global optimality of the iterate generated by either Bi-RRM or Bi-SGD. For Bi-RRM, we can employ the existing bilevel optimization solver as an oracle at each step given $\mathcal{D}_x(y^*(x_r))$ and $\mathcal{D}_y(x_r)$.

Results on the Synthetic Data Set. The results are shown in Figure 1 based on 10 independent runs. From Figure 1(a), it can be observed that as the value of ε increases from 1×10^{-3} to 0.1, the convergence rate of the Bi-RRM algorithm slows down, which aligns with our theoretical analysis. According to our theory, once ε surpasses a certain threshold, Bi-RRM fails to converge. The numerical results confirm this, as it is evident that when $\varepsilon = 1$ or $\varepsilon = 10$, Bi-RRM fails to find any stationary solution and may exhibit oscillatory behavior.

Another noteworthy characteristic of the bilevel problem is the asymmetry between the leader and the follower. As observed in Figure 1(b), the convergence rate of the Bi-RRM algorithm is more sensitive to the UL problem compared to the LL problem in terms of ε_x-sensitivity and ε_y-sensitivity. A notable comparison is the case of $\varepsilon_x = 100, \varepsilon_y = 0.1$ versus $\varepsilon_x = 0.1, \varepsilon_y = 100$. It is evident that the Bi-RRM algorithm converges to the global optimal solution of the problem when $\varepsilon_x = 0.1, \varepsilon_y = 100$, but fails to do so when $\varepsilon_x = 100, \varepsilon_y = 0.1$. This implies that the contribution of the LL optimization process (or the



(a) Bi-RRM ($\varepsilon_x = \varepsilon_y$) (b) Bi-RRM ($\varepsilon_x \neq \varepsilon_y$) (c) Bi-SGD ($\varepsilon_x = \varepsilon_y$)

Figure 1. Performance of Bi-RRM and Bi-SGD on the least-squares over different levels of the ε-sensitivity parameters.

follower) to the convergence of the entire sequence is not as sensitive as the UL optimization process (or the leader), which is again consistent with our theoretical result shown in Theorem 1.

7.2. Meta Strategic Classification

We also test the Bi-RRM and Bi-SGD on the problem of meta strategic classification, which can be simply written as follows.

$$\min_{x\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{Z_i\sim\mathcal{D}_i(y_i^*(x))} f_i(y_i^*(x); Z_i) + \frac{\lambda_x}{2}\|x\|^2$$
$$\text{s.t. } y_i^*(x) = \arg\min_{y_i\in\mathbb{R}^d} \mathop{\mathbb{E}}_{Z_i\sim\mathcal{D}_i(x)} \ell_i(x, y_i; Z_i), i \in [m]$$

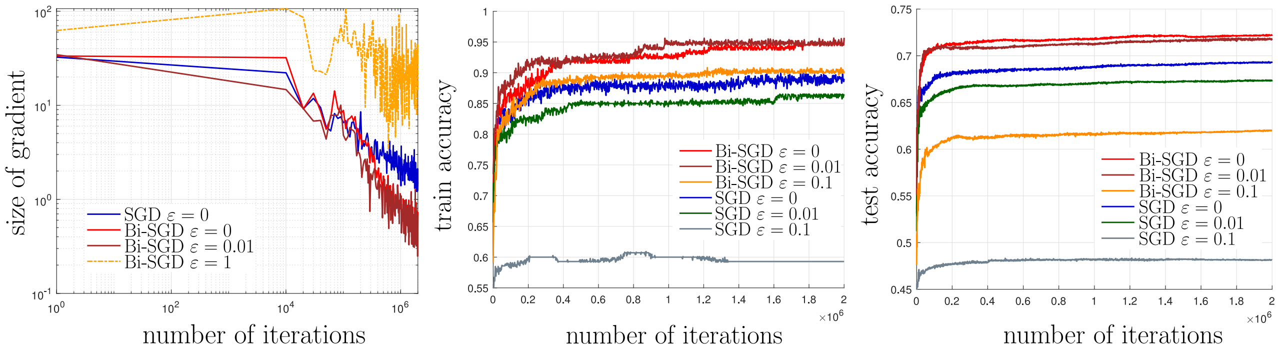
where $\ell_i(x, y_i; Z_i) \triangleq \langle y_i, \nabla f_i(x; Z_i)\rangle + \lambda_y/2\|y_i - x\|^2$, $\lambda_x, \lambda_y > 0$, and m denotes the total number of tasks. It is obvious that when $\mathcal{D}_i(y_i^*(x)) = \mathcal{D}_i(x) = D_i, \forall i$, this problem reduces to the classic formulation of MAML (Fallah et al., 2020; Finn et al., 2017). Here, the logistic regression loss is chosen as the objective function $f_i(;)$.

Results on the Spambase Data Set. In the numerical experiments, we employ the UCI machine learning repository spambase data set (Hopkins et al., 1999) for binary classification. This data set consists of 57 attributes with both continuous and discrete values, comprising a total of 4601 instances. Each instance is associated with a class label, denoted by 1 or 0, indicating whether it is categorized as spam or not. To demonstrate the effectiveness of meta-learning on this data set, we partition it into multiple subsets and treat each subset as an individual task. For each task, we create different data distributions as follows: initially, we randomly shuffle the entire data set and select 5 data samples for each task, resulting in a total of 5 tasks. Subsequently, we remove certain features from each task to create the training data set. More specifically, we set the ith and the $i+3$th columns of the data set to 0 for these 5 tasks, where i ranges from 1 to 5. Finally, we utilize 800 data samples that are not included in the training data set, with 5 samples used for meta-training and the remaining samples used for meta-testing. Overall, each task only has a subset of the spambase data set features, and the learned model is evaluated on the meta-testing data set that partially overlaps with the training data. This setup ensures that the features in the latent space are transferable, enabling us to assess the generalization performance of the learned model.

We conduct the numerical experiments to compare the performance of both Bi-SGD and traditional SGD (without PP training). We set the values of λ_x, λ_y to 1×10^{-3} and 1 respectively, and the step size is chosen as $0.5/\sqrt{10+r}$ for both Bi-SGD and SGD. The results, averaged over 10 independent trials, are presented in Figure 2. Figure 2(a) illustrates that when ε is large, Bi-SGD may not converge to a stationary point, as observed in the case when $\varepsilon = 1$. We further analyze the meta-training and meta-testing performance of both algorithms under different levels of sensitivity parameters. Figure 2(b) demonstrates that Bi-SGD, trained on the strategic data generating process, can perform well in the few-shot learning setting, yielding high meta-training accuracy.

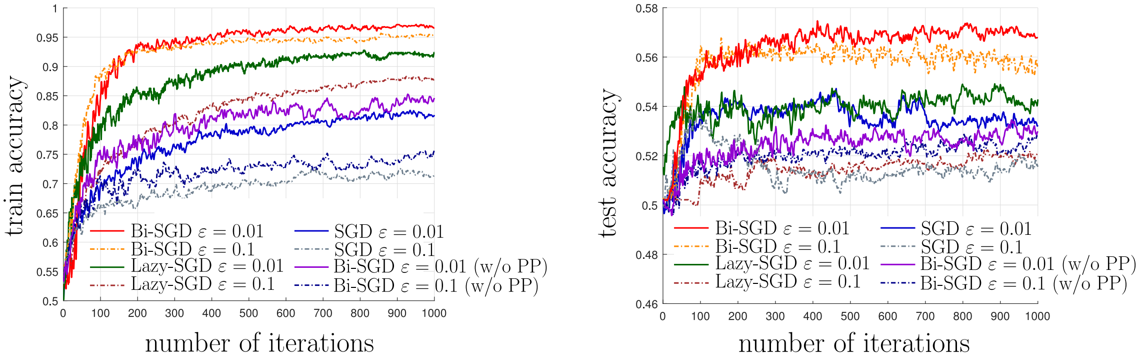
It is important to note that SGD is only trained on non-strategic features. In this case, the use of SGD may lead to reduced training accuracy and increased sensitivity to the ε-sensitivity parameters. The meta-testing results provide clearer insights into these characteristics. Firstly, when ε is small, the testing accuracy achieved by Bi-SGD is close to the maximum one, aligning with our theoretical analysis presented in Theorem 3. Secondly, the convergence rate and behavior of both Bi-SGD and SGD strongly depend on the levels of ε-sensitivity, emphasizing the significance of quantifying the maximum budget for the sensitivity parameters. Lastly, the meta-learning approach yields significantly higher testing accuracy compared to the single-level learning strategy.

Results on the Amazon Review Data Set. We also perform the numerical experiments using the UCI sentiment labeled sentences data set (Kotzias et al., 2015), specifically the Amazon reviews subset, which is a prominent area of research in natural language processing. To vectorize the sentences, we utilize the CountVectorizer from the scikit-learn library, generating word count vectors



(a) Stationarity (b) Train accuracy (c) Test accuracy

Figure 2. Performance comparison of Bi-SGD and SGD for meta strategic learning over different levels of the ε-sensitivity parameters on the spambase data set, where $\varepsilon \triangleq \varepsilon_x = \varepsilon_y$.



(a) Train accuracy (b) Test accuracy

Figure 3. Performance comparison of Bi-SGD, Bi-SGD without PP training, Lazy-SGD (aka Lazy deploy), and SGD without PP training for meta strategic learning over different levels of the ε-sensitivity parameters on the Amazon review data set, where $\varepsilon \triangleq \varepsilon_x = \varepsilon_y$.

for each sentence. Each data sample is represented by a vector of size 1546. For our experiments, we select 10 data samples for each of the 5 tasks, resulting in a total of 50 data samples. As the vectors are sparse, we partition each of them as 5 parts and set the entries of the ith part as zero for the ith task, creating heterogeneously distributed and partially observed data vectors for these tasks. Additionally, we use another set of 10 and 250 data samples as the meta-training and meta-testing data sets, respectively. We compare the performance of our proposed Bi-SGD algorithm with Lazy-SGD (aka Lazy deploy) (Mendler-Dünner et al., 2020), traditional SGD and bilevel SGD without PP training. The step sizes for the tested algorithms are chosen as $10/\sqrt{10+r}$.

The results, depicted in Figure 3, reveal that Bi-SGD achieves higher meta-training and meta-testing accuracies compared to the other two benchmark algorithms. This can be attributed to the bilevel structure, which excels at learning the invariant latent feature space. As a result, it leads to improved generalization during the meta-testing phase. Furthermore, we observe that increasing the value of ε corresponds to a decrease in accuracy, indicating that a larger deviation can result in a slower convergence rate, aligning with our theoretical findings. Additional numerical results can be found in Section E of the appendix.

8. Concluding Remarks

In this work, our focus is on stochastic bilevel optimization with applications to meta strategic learning. We begin by verifying the existence of the BPS point based on the ε-sensitivity parameters when strategic learners participate in the bilevel PP process. We then revisit the commonly used bilevel SGD method to solve this class of bilevel optimization problems in an iterative way. Moreover, we establish the convergence rate of Bi-SGD, which matches the standard rate of SGD for solving non-performative prediction problems. This convergence rate is achieved under the satisfaction of the newly provided existence conditions for the BPS point. Besides, we quantify an upper bound on the mismatch between the BPS and BPO points, demonstrating that a smaller ε-sensitivity parameter leads to closer proximity between these two points. The numerical experiments support and validate our theoretical results, providing empirical evidence for the effectiveness of our proposed approaches.

References

Balcan, M.-F., Khodak, M., and Talwalkar, A. Provable guarantees for gradient-based meta-learning. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 424–433, 2019.

Brown, G., Hod, S., and Kalemaj, I. Performative prediction in a stateful world. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 6045–6061, 2022.

Bubeck, S. et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Chen, L., Lu, S., and Chen, T. Understanding benign overfitting in gradient-based meta learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Chen, T., Sun, Y., and Yin, W. Tighter analysis of alternating stochastic gradient method for stochastic nested problems. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Chua, K., Lei, Q., and Lee, J. D. How fine-tuning allows for effective meta-learning. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 34: 8871–8884, 2021.

Cutler, J., Drusvyatskiy, D., and Harchaoui, Z. Stochastic optimization under distributional drift. *arXiv preprint arXiv:2108.07356*, 2021.

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

Drusvyatskiy, D. and Xiao, L. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 2022.

Fallah, A., Mokhtari, A., and Ozdaglar, A. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1082–1092, 2020.

Fiez, T., Chasnov, B., and Ratliff, L. Implicit learning dynamics in stackelberg games: Equilibria characterization, convergence analysis, and empirical study. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 3133–3144, 2020.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 1126–1135, 2017.

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

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

Hardt, M., Jagadeesan, M., and Mendler-Dünner, C. Performative power. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Hopkins, M., Reeber, E., Forman, G., and Suermondt, J. Spambase. *UCI Machine Learning Repository*, 1999.

Hu, Q., Zhong, Y., and Yang, T. Multi-block min-max bilevel optimization with applications in multi-task deep auc maximization. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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

Izzo, Z., Ying, L., and Zou, J. How to learn when data reacts to your model: performative gradient descent. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 4641–4650, 2021.

Izzo, Z., Zou, J., and Ying, L. How to learn when data gradually reacts to your model. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3998–4035, 2022.

Jagadeesan, M., Zrnic, T., and Mendler-Dünner, C. Regret minimization with performative feedback. *arXiv preprint arXiv:2202.00628*, 2022.

Ji, K., Yang, J., and Liang, Y. Bilevel optimization: Convergence analysis and enhanced design. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 4882–4892, 2021.

Kim, M. P. and Perdomo, J. C. Making decisions under outcome performativity. *arXiv preprint arXiv:2210.01745*, 2022.

Kotzias, D., Denil, M., De Freitas, N., and Smyth, P. From group to individual labels using deep features. In *Proceedings of ACM SIGKDD International Conference* *on Knowledge Discovery and Data Mining (KDD)*, pp. 597–606, 2015.

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

Li, Q. and Wai, H.-T. State dependent performative prediction with stochastic approximation. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Li, Q., Yau, C.-Y., and Wai, H.-T. Multi-agent performative prediction with greedy deployment and consensus seeking agents. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022b.

Liu, H., Socher, R., and Xiong, C. Taming MAML: Efficient unbiased meta-reinforcement learning. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 4061–4071, 2019.

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

Liu, R., Mu, P., Yuan, X., Zeng, S., and Zhang, J. A general descent aggregation framework for gradient-based bi-level optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Liu, Y., Chen, Y., Tang, Z., and Zhang, K. Model transferability with responsive decision subjects. *arXiv preprint arXiv:2107.05911*, 2021b.

Lu, S. and Gao, T. Meta-DAG: Meta causal discovery via bilevel optimization. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Maheshwari, C., Chiu, C.-Y., Mazumdar, E., Sastry, S., and Ratliff, L. Zeroth-order methods for convex-concave min-max problems: Applications to decision-dependent risk minimization. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 6702–6734, 2022.

Maity, S., Mukherjee, D., Banerjee, M., and Sun, Y. Predictor-corrector algorithms for stochastic optimization under gradual distribution shift. *arXiv preprint arXiv:2205.13575*, 2022.

Mandal, D., Triantafyllou, S., and Radanovic, G. Performative reinforcement learning. *arXiv preprint arXiv:2207.00046*, 2022.

Mendler-Dünner, C., Perdomo, J., Zrnic, T., and Hardt, M. Stochastic optimization for performative prediction. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 4929–4939, 2020.

Mendler-Dünner, C., Ding, F., and Wang, Y. Anticipating performativity by predicting from predictions. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Miller, J. P., Perdomo, J. C., and Zrnic, T. Outside the echo chamber: Optimizing the performative risk. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 7710–7720, 2021.

Mofakhami, M., Mitliagkas, I., and Gidel, G. Performative prediction with neural networks. In *NeurIPS Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022.

Narang, A., Faulkner, E., Drusvyatskiy, D., Fazel, M., and Ratliff, L. Learning in stochastic monotone games with decision-dependent data. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 5891–5912, 2022.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. Performative prediction. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 7599–7609, 13–18 Jul. 2020.

Rajeswaran, A., Finn, C., Kakade, S. M., and Levine, S. Meta-learning with implicit gradients. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Robbins, H. and Siegmund, D. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pp. 233–257. 1971.

Roy, A., Balasubramanian, K., and Ghadimi, S. Projection-free constrained stochastic nonconvex optimization with state-dependent markov data. *arXiv preprint arXiv:2206.11346*, 2022.

Sabach, S. and Shtern, S. A first order method for solving convex bilevel optimization problems. *SIAM Journal on Optimization*, 27(2):640–660, 2017.

Shen, H. and Chen, T. A single-timescale analysis for stochastic approximation with multiple coupled sequences. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

T Dinh, C., Tran, N., and Nguyen, J. Personalized federated learning with moreau envelopes. In *Proceedings* *of Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 21394–21405, 2020.

Wood, K. and DallAnese, E. Stochastic saddle point problems with decision-dependent distributions. *arXiv preprint arXiv:2201.02313*, 2022.

Wood, K., Bianchin, G., and DallAnese, E. Online projected gradient descent for stochastic optimization with decision-dependent distributions. *IEEE Control Systems Letters*, 6:1646–1651, 2021.

Ye, M., Liu, B., Wright, S., Stone, P., and Liu, Q. BOME! bilevel optimization made easy: A simple first-order approach. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Zhao, Y. Optimizing the performative risk under weak convexity assumptions. *arXiv preprint arXiv:2209.00771*, 2022.

Zrnic, T., Mazumdar, E., Sastry, S., and Jordan, M. Who leads and who follows in strategic classification? *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 15257–15269, 2021.

A. Preliminaries

Before showing the detailed derivations of the lemmas and theorems, we give the following inequalities and equalities which are often used in the proofs.

A.1. Inequalities

1. (Corollary 3.1 (Drusvyatskiy & Xiao, 2022)) Suppose that $f(y;Z)$ is C^1 smooth and the map $Z \mapsto \nabla f(y;Z)$ is L-Lipschitz continuous. Also, assume that there exists a $\varepsilon > 0$ satisfying $\mathcal{W}_1(\mathcal{D}(x), \mathcal{D}(x')) \le \varepsilon \|x - x'\|$. Then,

$$\sup_y \left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(x)} \nabla f(y;Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(x')} \nabla f(y;Z) \right\| \le \varepsilon L \left\| x - x' \right\|. \tag{23}$$

2. (Lemma 3.11 (Bubeck et al., 2015)). Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and γ-strongly convex, then for all $x, y \in \mathbb{R}^d$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \ge \frac{\gamma L}{\gamma + L} \|x - y\|^2 + \frac{1}{\gamma + L} \|\nabla f(x) - \nabla f(y)\|^2. \tag{24}$$

3. Young's inequality with parameter $\theta > 0$ is

$$\langle x, y \rangle \le \frac{1}{2\theta} \|x\|^2 + \frac{\theta}{2} \|y\|^2, \quad \forall x, y. \tag{25}$$

From the optimality condition of the LL problem, we have

$$\nabla^2_{xy} \ell(x, y^*(x)) + \nabla_x y^*(x)^T \nabla^2_{yy} \ell(x, y^*(x)) = 0, \tag{26}$$

which gives the gradient of the UL objective function as

$$\begin{aligned} \nabla_x F(x) = \nabla f(x, y^*(x)) &\triangleq \overline{\nabla} f(x, y^*(x)) \\ &= \nabla_x f(x, y^*(x)) - \nabla^2_{xy} \ell(x, y^*(x)) \left[\nabla^2_{yy} \ell(x, y^*(x)) \right]^{-1} \nabla_y f(x, y^*(x)) \end{aligned} \tag{27}$$

by applying the chain rule. By following this notation, the expression of $\overline{\nabla} f(x, y)$ is defined as follows (Ghadimi & Wang, 2018; Shen & Chen, 2022)

$$\overline{\nabla} f(x, y) = \nabla_x f(x, y) - \nabla^2_{xy} \ell(x, y) \left[\nabla^2_{yy} \ell(x, y) \right]^{-1} \nabla_y f(x, y). \tag{28}$$

Based on the definitions of $y^*(\varphi, x)$ shown in (5b) and $y^*(x)$ shown in (1b), we have

$$y^*(x, x) = y^*(x). \tag{29}$$

A.2. Lipschitz Constants

In this section, we present the Lipschitz constants used to quantify the changes in gradients. First, the detailed definitions of the Lipschitz continuous constants in Assumption 5 are further given as follows.

(Smoothness of the UL loss function) Assume that the gradient of loss function $f(x, y; Z)$ w.r.t. x is Lipschitz continuous $\forall x, x', y, y'$, namely,

$$\left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\cdot)} \nabla_x f(x, y; Z) - \nabla_x f(x', y; Z) \right\| \le L^x_f \left\| x - x' \right\|, \tag{30a}$$

$$\left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\cdot)} \nabla_x f(x, y; Z) - \nabla_x f(x, y'; Z) \right\| \le \bar{L}^x_f \left\| y - y' \right\|. \tag{30b}$$

Also, we assume that the gradient of $f(x, y; Z)$ w.r.t. y is Lipschitz continuous $\forall x, x', y, y'$, namely,

$$\left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\cdot)} \nabla_y f(x, y; Z) - \nabla_y f(x, y'; Z) \right\| \le L_f^y \|y - y'\|, \tag{31a}$$

$$\left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\cdot)} \nabla_y f(x, y; Z) - \nabla_y f(x, y'; Z) \right\| \le \bar{L}_f^y \|x - x'\|. \tag{31b}$$

(Smoothness of the LL loss function) Similarly, we also assume that loss function $\ell(x, y; Z)$ is L_ℓ^y-smooth, $\forall x, y, y'$ namely,

$$\left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\cdot)} \nabla_y \ell(x, y; Z) - \nabla_y \ell(x, y'; Z) \right\| \le L_\ell^y \|y - y'\|. \tag{32}$$

Assume that the Jacobian and Hessian matrices of loss function $\ell(x, y)$ are Lipschitz continuous $\forall x, x', y, y'$, namely,

$$\left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\cdot)} \nabla_{xy}^2 \ell(x, y; Z) - \nabla_{xy}^2 \ell(x', y; Z) \right\| \le L_{\ell xy}^x \left\| x - x' \right\|, \tag{33a}$$

$$\left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\cdot)} \nabla_{xy}^2 \ell(x, y; Z) - \nabla_{xy}^2 \ell(x, y'; Z) \right\| \le L_{\ell xy}^y \left\| y - y' \right\|, \tag{33b}$$

$$\left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\cdot)} \nabla_{yy}^2 \ell(x, y; Z) - \nabla_{yy}^2 \ell(x', y; Z) \right\| \le L_{\ell yy}^x \left\| x - x' \right\|, \tag{33c}$$

$$\left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\cdot)} \nabla_{yy}^2 \ell(x, y; Z) - \nabla_{yy}^2 \ell(x, y'; Z) \right\| \le L_{\ell yy}^y \left\| y - y' \right\|. \tag{33d}$$

Next, we will introduce the technical lemmas that quantify the changes in gradients with respect to shifts in the data distribution. These lemmas are used in the convergence proofs of both Bi-RRM and Bi-SGD. The following table provides a summary of the gradient Lipschitz continuity with respect to the decision-dependent distributions.

Table 2. Notations for Constants

Constant (abbrv.)	Definition	Details
$L_f^{\varepsilon_x}$	$\|\overline{\nabla} f(x, y) - \nabla f(x, y^*(x))\| \le L_f^{\varepsilon_x} \|y - y^*(x)\|$	Lemma 1
$L_y^{\varepsilon_y}$	$\|y^*(x) - y^*(x')\|^2 \le L_y^{\varepsilon_y} \|x - x'\|$	Lemma 2
$L_F^{\varepsilon_x,\varepsilon_y}$ (L_F)	$\|\overline{\nabla} f(x, y^*(x)) - \overline{\nabla} f(x', y^*(x'))\| \le L_F^{\varepsilon_x,\varepsilon_y} \|x - x'\|$	Lemma 3
$L_Z^{\varepsilon_x,\varepsilon_y}$ (L_Z)	$\left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(\varphi,x))} \overline{\nabla} f(x, y^*(\varphi, x); Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(\varphi,x'))} \overline{\nabla} f(x, y^*(\varphi, x'); Z) \right\| \le L_Z^{\varepsilon_x,\varepsilon_y} \|x - x'\|$	Lemma 4
$L_Z'^{\varepsilon_x,\varepsilon_y}$ (L_Z')	$\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x))} \overline{\nabla} f(x, y^*(x); Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(y^*(x'))} \overline{\nabla} f(x, y^*(x, x'); Z)\| \le L_Z'^{\varepsilon_x,\varepsilon_y} \|x - x'\|$	Lemma 4

Lemma 1. *The gradient of the UL objective function is Lipschitz continuous with constant $L_f^{\varepsilon_x}$, namely,*

$$\left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y)} \overline{\nabla} f(x, y; Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x))} \nabla f(x, y^*(x); Z) \right\| \le L_f^{\varepsilon_x} \|y - y^*(x)\| \tag{34}$$

where constant

$$L_f^{\varepsilon_x} \triangleq L_f^z \varepsilon_x + L_f^x + \frac{C_f^y L_{\ell xy}^y}{\gamma_y} + \frac{C_f^y C_{\ell xy} L_{\ell yy}^y}{\gamma_y^2} + \frac{C_{\ell xy}}{\gamma_y} \left(L_f^z \varepsilon_x + L_f^y \right). \tag{35}$$

Proof. From (27) and (28), we can have

$$\begin{align}
&\left\|\overline{\nabla} f(x,y) - \nabla f(x,y^*(x))\right\| \nonumber\\
\le{}& \Big\|\nabla_x f(x,y) - \nabla^2_{xy}\ell(x,y)\left[\nabla^2_{yy}\ell(x,y)\right]^{-1}\nabla_y f(x,y) \nonumber\\
&- \left(\nabla_x f(x,y^*(x)) - \nabla^2_{xy}\ell(x,y^*(x))\left[\nabla^2_{yy}\ell(x,y^*(x))\right]^{-1}\nabla_y f(x,y^*(x))\right)\Big\| \tag{36}\\
\le{}& \left\|\underset{Z\sim\mathcal{D}_x(y)}{\mathbb{E}}\nabla_x f(x,y;Z) - \underset{Z\sim\mathcal{D}_x(y^*(x))}{\mathbb{E}}\nabla_x f(x,y^*(x);Z)\right\| \nonumber\\
&+ \Bigg\|\underset{Z\sim\mathcal{D}_y(x)}{\mathbb{E}}\nabla^2_{xy}\ell(x,y;Z)\left[\underset{Z\sim\mathcal{D}_y(x)}{\mathbb{E}}\nabla^2_{yy}\ell(x,y;Z)\right]^{-1}\underset{Z\sim\mathcal{D}_x(y)}{\mathbb{E}}\nabla_y f(x,y;Z) \nonumber\\
&- \underset{Z\sim\mathcal{D}_y(x)}{\mathbb{E}}\nabla^2_{xy}\ell(x,y^*(x);Z)\left[\underset{Z\sim\mathcal{D}_y(x)}{\mathbb{E}}\nabla^2_{yy}\ell(x,y^*(x);Z)\right]^{-1}\underset{Z\sim\mathcal{D}_x(y^*(x))}{\mathbb{E}}\nabla_y f(x,y^*(x);Z)\Bigg\| \tag{37}\\
\le{}& \left\|\underset{Z\sim\mathcal{D}_x(y)}{\mathbb{E}}\nabla_x f(x,y;Z) - \underset{Z\sim\mathcal{D}_x(y^*(x))}{\mathbb{E}}\nabla_x f(x,y;Z)\right\| + \left\|\underset{Z\sim\mathcal{D}_x(y^*(x))}{\mathbb{E}}\nabla_x f(x,y;Z) - \underset{Z\sim\mathcal{D}_x(y^*(x))}{\mathbb{E}}\nabla_x f(x,y^*(x);Z)\right\| \nonumber\\
&+ \|\Delta_1\| + \|\Delta_2\| + \|\Delta_3\| \tag{38}\\
\overset{(23)}{\le}{}& (L^z_f\varepsilon_x + \bar{L}^x_f)\|y-y^*(x)\| + \|\Delta_1\| + \|\Delta_2\| + \|\Delta_3\| \tag{39}
\end{align}$$

where

$$\Delta_1 \triangleq \left(\underset{Z\sim\mathcal{D}_y(x)}{\mathbb{E}}\nabla^2_{xy}\ell(x,y;Z) - \underset{Z\sim\mathcal{D}_y(x)}{\mathbb{E}}\nabla^2_{xy}\ell(x,y^*(x);Z)\right)\left[\nabla^2_{yy}\ell(x,y)\right]^{-1}\nabla_y f(x,y), \tag{40a}$$

$$\Delta_2 \triangleq \nabla^2_{xy}\ell(x,y^*(x))\left(\left[\underset{Z\sim\mathcal{D}_y(x)}{\mathbb{E}}\nabla^2_{yy}\ell(x,y;Z)\right]^{-1} - \left[\underset{Z\sim\mathcal{D}_y(x)}{\mathbb{E}}\nabla^2_{yy}\ell(x,y^*(x);Z)\right]^{-1}\right)\nabla_y f(x,y), \tag{40b}$$

$$\Delta_3 \triangleq \nabla^2_{xy}\ell(x,y^*(x))\left[\nabla^2_{yy}\ell(x,y^*(x))\right]^{-1}\left(\underset{Z\sim\mathcal{D}_x(y)}{\mathbb{E}}\nabla_y f(x,y;Z) - \underset{Z\sim\mathcal{D}_x(y^*(x))}{\mathbb{E}}\nabla_y f(x,y^*(x);Z)\right). \tag{40c}$$

Note that

$$\|\Delta_1\| \le \left\|\left(\underset{Z\sim\mathcal{D}_y(x)}{\mathbb{E}}\nabla^2_{xy}\ell(x,y;Z) - \underset{Z\sim\mathcal{D}_y(x)}{\mathbb{E}}\nabla^2_{xy}\ell(x,y^*(x);Z)\right)\left[\nabla^2_{yy}\ell(x,y)\right]^{-1}\nabla_y f(x,y)\right\| \le \frac{C^y_f L^y_{\ell xy}}{\gamma_y}\|y-y^*(x)\|,$$

and similarly we can also obtain

$$\begin{align}
\|\Delta_2\| \le{}& \left\|\nabla^2_{xy}\ell(x,y^*(x))\left(\left[\underset{Z\sim\mathcal{D}_y(x)}{\mathbb{E}}\nabla^2_{yy}\ell(x,y;Z)\right]^{-1} - \left[\underset{Z\sim\mathcal{D}_y(x)}{\mathbb{E}}\nabla^2_{yy}\ell(x,y^*(x);Z)\right]^{-1}\right)\nabla_y f(x,y)\right\| \nonumber\\
\le{}& \frac{C_{\ell xy}C^y_f L^y_{\ell yy}}{\gamma_y^2}\|y-y^*(x)\| \tag{41}
\end{align}$$

where we use the fact $\|H_2^{-1} - H_1^{-1}\| = \|H_1^{-1}(H_1 - H_2)H_2^{-1}\| \le \|H_1^{-1}\|\|H_2^{-1}\|\|H_1 - H_2\|$, and

$$\begin{align}
\|\Delta_3\| \le{}& \left\|\nabla^2_{xy}\ell(x,y^*(x))\left[\nabla^2_{yy}\ell(x,y^*(x))\right]^{-1}\left(\underset{Z\sim\mathcal{D}_x(y)}{\mathbb{E}}\nabla_y f(x,y;Z) - \underset{Z\sim\mathcal{D}_x(y^*(x))}{\mathbb{E}}\nabla_y f(x,y;Z)\right)\right\| \nonumber\\
&+ \left\|\nabla^2_{xy}\ell(x,y^*(x))\left[\nabla^2_{yy}\ell(x,y^*(x))\right]^{-1}\left(\underset{Z\sim\mathcal{D}_x(y^*(x))}{\mathbb{E}}\nabla_y f(x,y;Z) - \underset{Z\sim\mathcal{D}_x(y^*(x))}{\mathbb{E}}\nabla_y f(x,y^*(x);Z)\right)\right\| \nonumber\\
\overset{(23)}{\le}{}& \frac{C_{\ell xy}}{\gamma_y}\left(L^z_f\varepsilon_x + L^y_f\right)\|y-y^*(x)\|. \tag{42}
\end{align}$$

Therefore, we have

$$\begin{aligned}&\left\|\bar{\nabla} f(x,y)-\nabla f(x,y^*(x))\right\|\\ \le&\underbrace{L_f^z\varepsilon_x+\bar{L}_f^x+\frac{C_f^y L_{\ell xy}^y}{\gamma_y}+\frac{C_f^y C_{\ell xy}L_{\ell yy}^y}{\gamma_y^2}+\frac{C_{\ell xy}}{\gamma_y}\left(L_f^z\varepsilon_x+L_f^y\right)}_{\triangleq L_f^{\varepsilon_x}}\|y-y^*(x)\|.\end{aligned}\tag{43}$$

□

Lemma 2. *Function $y^*(\varphi,x)$ is Lipschitz continuous (w.r.t. x) with constant $L_\ell^z\varepsilon_y\gamma_y^{-1}$, namely,*

$$\|y^*(\varphi,x)-y^*(\varphi,x')\|\le\frac{L_\ell^z\varepsilon_y}{\gamma_y}\|x-x'\|,\quad\forall\varphi.\tag{44}$$

Also, function $y^*(x)$ is Lipschitz continuous (w.r.t. x) with constant $L_y^{\varepsilon_y}$, namely,

$$\|y^*(x)-y^*(x')\|^2\le L_y^{\varepsilon_y}\|x-x'\|\tag{45}$$

where constant

$$L_y^{\varepsilon_y}\triangleq\frac{C_{\ell xy}+L_\ell^z\varepsilon_y}{\gamma_y}.\tag{46}$$

Proof. First, define the following auxiliary variables:

$$y^*(\varphi,x)=\arg\min_y\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x)}\ell(\varphi,y;Z),\tag{47a}$$

$$y^*(\varphi,x')=\arg\min_y\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x')}\ell(\varphi,y;Z).\tag{47b}$$

As $\ell(x,y;Z)$ is strongly convex, we can have

$$\left\langle\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x)}\nabla\ell(\varphi,y^*(\varphi,x);Z)-\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x)}\nabla\ell(\varphi,y^*(\varphi,x');Z),y^*(\varphi,x)-y^*(\varphi,x')\right\rangle\ge\gamma_y\|y^*(\varphi,x)-y^*(\varphi,x')\|^2.\tag{48}$$

From the optimality conditions of (47a) and (47b), we have $\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x')}\nabla_y\ell(\varphi,y^*(\varphi,x');Z)=0$ and $\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x)}\nabla_y\ell(\varphi,y^*(\varphi,x);Z)=0$. Therefore, we can replace $\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x)}\nabla\ell(\varphi,y^*(\varphi,x);Z)$ by $\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x')}\nabla\ell(\varphi,y^*(\varphi,x');Z)$ in (48). As a result, we obtain

$$\begin{aligned}&\gamma_y\|y^*(\varphi,x)-y^*(\varphi,x')\|^2\\ \le&\left\langle\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x')}\nabla\ell(\varphi,y^*(\varphi,x');Z)-\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x)}\nabla\ell(\varphi,y^*(\varphi,x');Z),y^*(\varphi,x)-y^*(\varphi,x')\right\rangle &(49)\\ \overset{(23)}{\le}&L_\ell^z\varepsilon_y\|x-x'\|\|y^*(\varphi,x)-y^*(\varphi,x')\|, &(50)\end{aligned}$$

which is equivalent to

$$\|y^*(\varphi,x)-y^*(\varphi,x')\|\le\frac{L_\ell^z\varepsilon_y}{\gamma_y}\|x-x'\|.\tag{51}$$

Recall

$$y^*(x)=\arg\min_y\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x)}\ell(x,y;Z),\tag{52}$$

so we know that

$$y^*(x')=\arg\min_y\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x')}\ell(x',y;Z).\tag{53}$$

From (26), we have

$$\|\nabla y^*(x)\| \leq \left\|\nabla^2_{xy}\ell(x, y^*(x))\right\| \left\|\left[\nabla^2_{yy}\ell(x, y^*(x))\right]^{-1}\right\| \leq \frac{C_{\ell xy}}{\gamma_y}, \tag{54}$$

which gives

$$\|y^*(x) - y^*(x', x)\| \leq \frac{C_{\ell xy}}{\gamma_y}\|x - x'\|, \tag{55}$$

as both $y^*(x)$ and $y^*(x', x)$ are obtained based on data sample $Z \sim \mathcal{D}_y(x)$.

Combining (55) and (51) yields

$$\begin{aligned}
\|y^*(x) - y^*(x')\| &\leq \|y^*(x) - y^*(x', x)\| + \|y^*(x', x) - y^*(x')\| &(56)\\
&\overset{(a)}{\leq} \left(\frac{C_{\ell xy}}{\gamma_y} + \frac{L^z_\ell \varepsilon_y}{\gamma_y}\right)\|x - x'\| &(57)\\
&\leq \underbrace{\frac{C_{\ell xy} + L^z_\ell \varepsilon_y}{\gamma_y}}_{\triangleq L^{\varepsilon_y}_y}\|x - x'\|. &(58)
\end{aligned}$$

where in (a) we substitute φ into (51) with x'. □

Lemma 3. *The gradient of the UL objective function is Lipschitz continuous (w.r.t. x) with constant $L_F^{\varepsilon_x,\varepsilon_y}$ (abbreviated as L_F), namely,*

$$\|\nabla_x F(x) - \nabla_x F(x')\| = \|\overline{\nabla} f(x, y^*(x)) - \overline{\nabla} f(x', y^*(x'))\| \leq L_F^{\varepsilon_x,\varepsilon_y}\|x - x'\| \tag{59}$$

where constant

$$\begin{aligned}
L_F^{\varepsilon_x,\varepsilon_y} \triangleq L^{\varepsilon_y}_y\left(L^z_f\varepsilon_x + \bar{L}^x_f\right) + L^x_f + \frac{\left(L^z_{\ell xy}\varepsilon_y + L^x_{\ell xy} + L^y_{\ell xy}L^{\varepsilon_y}_y\right)C^y_f}{\gamma_y}&\\
+ \frac{C_{\ell xy}C^y_f\left(L^z_{\ell yy}\varepsilon_y + L^x_{\ell yy} + L^{\varepsilon_y}_y L^y_{\ell yy}\right)}{\gamma_y^2} + \frac{C_{\ell xy}\left(L^{\varepsilon_y}_y L^z_f\varepsilon_x + \bar{L}^y_f + L^{\varepsilon_y}_y L^y_f\right)}{\gamma_y}.& \qquad (60)
\end{aligned}$$

Proof. According to the definition of $\overline{\nabla} f(x, y^*(x))$, we can have

$$\begin{aligned}
&\|\nabla F(x) - \nabla F(x')\|\\
&\leq \Bigg\|\nabla_x f(x, y^*(x)) - \nabla^2_{xy}\ell(x, y^*(x))\left[\nabla^2_{yy}\ell(x, y^*(x))\right]^{-1}\nabla_y f(x, y^*(x))\\
&\qquad - (\nabla_x f(x', y^*(x')) - \nabla^2_{xy}\ell(x', y^*(x'))\left[\nabla^2_{yy}\ell(x', y^*(x'))\right]^{-1}\nabla_y f(x', y^*(x')))\Bigg\| \qquad (61)\\
&\overset{(a)}{\leq} \|\nabla_x f(x, y^*(x)) - \nabla_x f(x', y^*(x'))\|\\
&\quad + \left\|\nabla^2_{xy}\ell(x, y^*(x))\left[\nabla^2_{yy}\ell(x, y^*(x))\right]^{-1}\nabla_y f(x, y^*(x)) - \nabla^2_{xy}\ell(x', y^*(x'))\left[\nabla^2_{yy}\ell(x', y^*(x'))\right]^{-1}\nabla_y f(x', y^*(x')))\right\|
\end{aligned}$$

where (a) is true due to the triangle inequality.

Next, let us define the following quantities:

$$\Delta_1 \triangleq \left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x)}\nabla^2_{xy}\ell(x, y^*(x); Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x')}\nabla^2_{xy}\ell(x', y^*(x'); Z)\right)\left[\nabla^2_{yy}\ell(x, y^*(x))\right]^{-1}\nabla_y f(x, y^*(x)), \tag{62a}$$

$$\Delta_2 \triangleq \nabla^2_{xy}\ell(x', y^*(x'))\left(\left[\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x)}\nabla^2_{yy}\ell(x, y^*(x); Z)\right]^{-1} - \left[\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x')}\nabla^2_{yy}\ell(x', y^*(x'); Z)\right]^{-1}\right)\nabla_y f(x, y^*(x)), \tag{62b}$$

$$\Delta_3 \triangleq \nabla^2_{xy}\ell(x', y^*(x'))\left[\nabla^2_{yy}\ell(x', y^*(x'))\right]^{-1}\left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}(y^*(x))}\nabla_y f(x, y^*(x); Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(y^*(x'))}\nabla_y f(x', y^*(x'); Z)\right). \tag{62c}$$

Then, we can have

$$\begin{aligned}
&\|\nabla F(x)-\nabla F(x')\| \\
\le\;& \|\nabla_x f(x,y^*(x))-\nabla_x f(x',y^*(x'))\|+\|\Delta_1\|+\|\Delta_2\|+\|\Delta_3\| \qquad (63)\\
\le\;& \Bigg(L_y^{\varepsilon_y}\left(L_f^z\varepsilon_x+\bar{L}_f^x\right)+L_f^x+\frac{\left(L_{\ell xy}^z\varepsilon_y+L_{\ell xy}^x+L_{\ell xy}^y L_y^{\varepsilon_y}\right)C_f^y}{\gamma_y} \\
&+\frac{C_{\ell xy}C_f^y\left(L_{\ell yy}^z\varepsilon_y+L_{\ell yy}^x+L_y^{\varepsilon_y}L_{\ell yy}^y\right)}{\gamma_y^2}+\frac{C_{\ell xy}\left(L_y^{\varepsilon_y}L_f^z\varepsilon_x+\bar{L}_f^y+L_y^{\varepsilon_y}L_f^y\right)}{\gamma_y}\Bigg)\|x-x'\|, \qquad (64)
\end{aligned}$$

where we use the following facts: first we have

$$\begin{aligned}
&\|\nabla_x f(x,y^*(x))-\nabla_x f(x',y^*(x'))\| \\
\le\;& \left\|\mathbb{E}_{Z\sim\mathcal{D}_x(y^*(x))}\nabla_x f(x,y^*(x))-\mathbb{E}_{Z\sim\mathcal{D}_x(y^*(x'))}\nabla_x f(x',y^*(x'))\right\| \qquad (65)\\
\le\;& \left\|\mathbb{E}_{Z\sim\mathcal{D}_x(y^*(x))}\nabla_x f(x,y^*(x))-\mathbb{E}_{Z\sim\mathcal{D}_x(y^*(x'))}\nabla_x f(x,y^*(x))\right\| \\
&+\left\|\mathbb{E}_{Z\sim\mathcal{D}_x(y^*(x'))}\nabla_x f(x,y^*(x))-\mathbb{E}_{Z\sim\mathcal{D}_x(y^*(x'))}\nabla_x f(x',y^*(x'))\right\| \qquad (66)\\
\overset{(23)}{\le}\;& \left(L_f^z\varepsilon_x+\bar{L}_f^x\right)\|y^*(x)-y^*(x')\|+L_f^x\|x-x'\| \qquad (67)\\
\overset{(45)}{\le}\;& \left(L_y^{\varepsilon_y}\left(L_f^z\varepsilon_x+\bar{L}_f^x\right)+L_f^x\right)\|x-x'\|, \qquad (68)
\end{aligned}$$

and similarly we can have

$$\begin{aligned}
&\left\|\mathbb{E}_{Z\sim\mathcal{D}_y(x)}\nabla_{xy}^2\ell(x,y^*(x);Z)-\mathbb{E}_{Z\sim\mathcal{D}_y(x')}\nabla_{xy}^2\ell(x',y^*(x');Z)\right\| \\
\overset{(23)}{\le}\;& \left(L_{\ell xy}^z\varepsilon_y+L_{\ell xy}^x\right)\|x-x'\|+L_{\ell xy}^y\|y^*(x)-y^*(x')\| \qquad (69)\\
\overset{(45)}{\le}\;& \left(L_{\ell xy}^z\varepsilon_y+L_{\ell xy}^x+L_{\ell xy}^y L_y^{\varepsilon_y}\right)\|x-x'\|, \qquad (70)
\end{aligned}$$

and

$$\left\|\mathbb{E}_{Z\sim\mathcal{D}_y(x)}\nabla_{yy}^2\ell(x,y^*(x);Z)-\mathbb{E}_{Z\sim\mathcal{D}_y(x')}\nabla_{yy}^2\ell(x',y^*(x');Z)\right\|\overset{(23),(45)}{\le}\left(L_{\ell yy}^z\varepsilon_y+L_{\ell yy}^x+L_y^{\varepsilon_y}L_{\ell yy}^y\right)\|x-x'\|, \qquad (71)$$

and

$$\left\|\mathbb{E}_{Z\sim\mathcal{D}(y^*(x))}\nabla_y f(x,y^*(x);Z)-\mathbb{E}_{Z\sim\mathcal{D}(y^*(x'))}\nabla_y f(x',y^*(x');Z)\right\|\overset{(23),(45)}{\le}\left(L_y^{\varepsilon_y}L_f^z\varepsilon_x+\bar{L}_f^y+L_y^{\varepsilon_y}L_f^y\right)\|x-x'\|. \qquad (72)$$

□

Lemma 4. *The gradient of the UL objective function is Lipschitz continuous (w.r.t. x) with constant $L_Z^{\varepsilon_x,\varepsilon_y}(L_Z),\forall\varphi,x,x'$, namely,*

$$\left\|\mathbb{E}_{Z\sim\mathcal{D}_x(y^*(\varphi,x))}\overline{\nabla}f(\varphi,y^*(\varphi,x);Z)-\mathbb{E}_{Z\sim\mathcal{D}_x(y^*(\varphi,x'))}\overline{\nabla}f(\varphi,y^*(\varphi,x');Z)\right\|\le L_Z^{\varepsilon_x,\varepsilon_y}\|x-x'\| \qquad (73)$$

where constant

$$L_Z^{\varepsilon_x,\varepsilon_y}\triangleq\left(\left(L_f^z+\frac{C_{\ell xy}L_f^z}{\gamma_y}\right)\varepsilon_x+\bar{L}_f^x+\frac{C_f^y L_{\ell xy}^y}{\gamma_y}+\frac{C_{\ell xy}}{\gamma_y}\left(L_f^y+\frac{C_f^y L_{\ell yy}^y}{\gamma_y}\right)\right)\frac{L_\ell^z\varepsilon_y}{\gamma_y}+\frac{C_f^y}{\gamma_y}\left(L_{\ell xy}^z+\frac{C_{\ell xy}L_{\ell yy}^z}{\gamma_y}\right)\varepsilon_y. \qquad (74)$$

Also, the gradient of the UL objective function is Lipschitz continuous (w.r.t. x) with constant $L'^{\varepsilon_x,\varepsilon_y}_Z (L'_Z), \forall x, x'$, *namely,*

$$\left\| \underset{Z\sim\mathcal{D}_x(y^*(x))}{\mathbb{E}} \overline{\nabla} f(x, y^*(x); Z) - \underset{Z\sim\mathcal{D}(y^*(x'))}{\mathbb{E}} \overline{\nabla} f(x, y^*(x, x'); Z) \right\| \leq L'^{\varepsilon_x,\varepsilon_y}_Z \|x - x'\|$$

where constant

$$L'^{\varepsilon_x,\varepsilon_y}_Z \triangleq L^{\varepsilon_x,\varepsilon_y}_Z + \left(L^z_f + \frac{C_{\ell xy} L^z_f}{\gamma_y} \right) \frac{C_{\ell xy}\varepsilon_x}{\gamma_y}. \tag{75}$$

Proof. Based on the closed form of $\overline{\nabla} f(x, y^*(\varphi, x))$, we have

$$\begin{aligned}
&\left\| \underset{Z\sim\mathcal{D}_x(y^*(\varphi,x))}{\mathbb{E}} \overline{\nabla} f(x, y^*(\varphi, x); Z) - \underset{Z\sim\mathcal{D}(y^*(\varphi,x'))}{\mathbb{E}} \overline{\nabla} f(x, y^*(\varphi, x'); Z) \right\| \\
\leq &\Bigg\| \underset{Z\sim\mathcal{D}_x(y^*(\varphi,x))}{\mathbb{E}} \nabla_x f(x, y^*(\varphi, x); Z) - \underset{Z\sim\mathcal{D}_x(y^*(\varphi,x'))}{\mathbb{E}} \nabla_x f(x, y^*(\varphi, x'); Z) \\
&\quad + \nabla^2_{xy}\ell(x, y^*(\varphi, x)) \left[\nabla^2_{yy}\ell(x, y^*(\varphi, x))\right]^{-1} \nabla_y f(x, y^*(\varphi, x)) \\
&\quad - \left(\nabla^2_{xy}\ell(x, y^*(\varphi, x')) \left[\nabla^2_{yy}\ell(x, y^*(\varphi, x'))\right]^{-1} \nabla_y f(x, y^*(\varphi, x')) \right) \Bigg\| \qquad (76) \\
\overset{(23)}{\leq} &\left(\bar{L}^x_f + L^z_f \varepsilon_x \right) \|y^*(\varphi, x) - y^*(\varphi, x')\| + \|\Delta_1\| + \|\Delta_2\| + \|\Delta_3\| \\
\overset{(a)}{\leq} &\left(\bar{L}^x_f + L^z_f\varepsilon_x + \frac{C^y_f L^y_{\ell xy}}{\gamma_y} + \frac{C_{\ell xy} C^y_f L^y_{\ell yy}}{\gamma_y^2} + \frac{C_{\ell xy}\left(L^z_f \varepsilon_x + L^y_f \right)}{\gamma_y} \right) \|y^*(\varphi, x) - y^*(\varphi, x')\| \\
&+ \frac{C^y_f}{\gamma_y} \left(L^z_{\ell xy} + \frac{C_{\ell xy} L^z_{\ell yy}}{\gamma_y} \right) \varepsilon_y \|x - x'\| \qquad (77)
\end{aligned}$$

where in (a) we use the following facts

$$\Delta_1 \triangleq \left(\underset{Z\sim\mathcal{D}_y(x)}{\mathbb{E}} \nabla^2_{xy}\ell(x, y^*(\varphi, x); Z) - \underset{Z\sim\mathcal{D}_y(x')}{\mathbb{E}} \nabla^2_{xy}\ell(x, y^*(\varphi, x); Z) \right) \left[\nabla^2_{yy}\ell(x, y^*(\varphi, x))\right]^{-1} \nabla_y f(x, y^*(\varphi, x)), \tag{78a}$$

$$\Delta_2 \triangleq \nabla^2_{xy}\ell(x, y^*(\varphi, x')) \left(\left[\underset{Z\sim\mathcal{D}_y(x)}{\mathbb{E}} \nabla^2_{yy}\ell(x, y^*(\varphi, x); Z) \right]^{-1} - \left[\underset{Z\sim\mathcal{D}_y(x')}{\mathbb{E}} \nabla^2_{yy}\ell(x, y^*(\varphi, x'); Z) \right]^{-1} \right) \nabla_y f(x, y^*(x)), \tag{78b}$$

$$\Delta_3 \triangleq \nabla^2_{xy}\ell(x, y^*(\varphi, x')) \left[\nabla^2_{yy}\ell(x, y^*(\varphi, x'))\right]^{-1} \left(\underset{Z\sim\mathcal{D}_x(y^*(\varphi,x))}{\mathbb{E}} \nabla_y f(x, y^*(\varphi, x); Z) - \underset{Z\sim\mathcal{D}_x(y^*(\varphi,x'))}{\mathbb{E}} \nabla_y f(x, y^*(\varphi, x'); Z) \right) \tag{78c}$$

with

$$\|\Delta_1\| \overset{(23)}{\leq} \frac{C^y_f L^z_{\ell xy} \varepsilon_y}{\gamma_y} \|x - x'\| + \frac{C^y_f L^y_{\ell xy}}{\gamma_y} \|y^*(\varphi, x) - y^*(\varphi, x')\|, \tag{79a}$$

$$\|\Delta_2\| \overset{(23)}{\leq} \frac{C_{\ell xy} C^y_f L^z_{\ell yy} \varepsilon_y}{\gamma_y^2} \|x - x'\| + \frac{C_{\ell xy} C^y_f L^y_{\ell yy}}{\gamma_y^2} \|y^*(\varphi, x) - y^*(\varphi, x')\|, \tag{79b}$$

$$\|\Delta_3\| \overset{(23)}{\leq} \frac{C_{\ell xy}\left(L^z_f \varepsilon_x + L^y_f \right)}{\gamma_y} \|y^*(\varphi, x) - y^*(\varphi, x')\|. \tag{79c}$$

Plugging (51) into (77) gives (73) and (74) immediately.

Similarly, we also have

$$
\begin{aligned}
&\left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x))} \overline{\nabla} f(x, y^*(x); Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x'))} \overline{\nabla} f(x, y^*(x, x'); Z) \right\| \\
\leq &\Bigg\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x))} \nabla_x f(x, y^*(x); Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x'))} \nabla_x f(x, y^*(x, x'); Z) \\
&\quad + \nabla^2_{xy} \ell(x, y^*(x)) \left[\nabla^2_{yy} \ell(x, y^*(x))\right]^{-1} \nabla_y f(x, y^*(x)) \\
&\quad - \left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x')} \nabla^2_{xy} \ell(x, y^*(x, x'); Z) \left[\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x')} \nabla^2_{yy} \ell(x, y^*(x, x'); Z) \right]^{-1} \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x'))} \nabla_y f(x, y^*(x, x'); Z) \right) \Bigg\| \qquad (80) \\
\overset{(23)}{\leq} &\bar{L}^x_f \|y^*(x) - y^*(x, x')\| + L^z_f \varepsilon_x \|y^*(x) - y^*(x')\| + \|\Delta'_1\| + \|\Delta'_2\| + \|\Delta'_3\| \\
\leq &\left(\bar{L}^x_f + \frac{C^y_f L^y_{\ell xy}}{\gamma_y} + \frac{C_{\ell xy} C^y_f L^y_{\ell yy}}{\gamma_y^2} + \frac{C_{\ell xy} L^y_f}{\gamma_y} \right) \|y^*(x) - y^*(x, x')\| \\
&+ \frac{C^y_f}{\gamma_y} \left(L^z_{\ell xy} + \frac{C_{\ell xy} L^z_{\ell yy}}{\gamma_y} \right) \varepsilon_y \|x - x'\| + \left(L^z_f + \frac{C_{\ell xy} L^z_f}{\gamma_y} \right) \varepsilon_x \|y^*(x) - y^*(x')\| \qquad (81)
\end{aligned}
$$

where

$$
\Delta'_1 \triangleq \left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x)} \nabla^2_{xy} \ell(x, y^*(x); Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x')} \nabla^2_{xy} \ell(x, y^*(x, x'); Z) \right) \left[\nabla^2_{yy} \ell(x, y^*(x))\right]^{-1} \nabla_y f(x, y^*(x)), \qquad (82a)
$$

$$
\begin{aligned}
\Delta'_2 \triangleq &\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x')} \nabla^2_{xy} \ell(x, y^*(x, x'); Z) \\
&\times \left(\left[\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x)} \nabla^2_{yy} \ell(x, y^*(x); Z) \right]^{-1} - \left[\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x')} \nabla^2_{yy} \ell(x, y^*(x, x'); Z) \right]^{-1} \right) \nabla_y f(x, y^*(x)), \qquad (82b)
\end{aligned}
$$

$$
\begin{aligned}
\Delta'_3 \triangleq &\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x')} \nabla^2_{xy} \ell(x, y^*(x, x'); Z) \left[\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x')} \nabla^2_{yy} \ell(x, y^*(x, x'); Z) \right]^{-1} \\
&\times \left(\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x))} \nabla_y f(x, y^*(x); Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x'))} \nabla_y f(x, y^*(x, x'); Z) \right) \qquad (82c)
\end{aligned}
$$

with

$$
\|\Delta'_1\| \overset{(23)}{\leq} \frac{C^y_f L^z_{\ell xy} \varepsilon_y}{\gamma_y} \|x - x'\| + \frac{C^y_f L^y_{\ell xy}}{\gamma_y} \|y^*(x) - y^*(x, x')\|, \qquad (83a)
$$

$$
\|\Delta'_2\| \overset{(23)}{\leq} \frac{C_{\ell xy} C^y_f L^z_{\ell yy} \varepsilon_y}{\gamma_y^2} \|x - x'\| + \frac{C_{\ell xy} C^y_f L^y_{\ell yy}}{\gamma_y^2} \|y^*(x) - y^*(x, x')\|, \qquad (83b)
$$

$$
\|\Delta'_3\| \overset{(23)}{\leq} \frac{C_{\ell xy} L^z_f \varepsilon_x}{\gamma_y} \|y^*(x) - y^*(x')\| + \frac{C_{\ell xy} L^y_f}{\gamma_y} \|y^*(x) - y^*(x, x')\|. \qquad (83c)
$$

Plugging (51) and (45) into (81) gives

$$
\begin{aligned}
&\left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x))} \overline{\nabla} f(x, y^*(x); Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x'))} \overline{\nabla} f(x, y^*(x, x'); Z) \right\| \\
\leq &\left(\bar{L}^x_f + \frac{C^y_f L^y_{\ell xy}}{\gamma_y} + \frac{C_{\ell xy} C^y_f L^y_{\ell yy}}{\gamma_y^2} + \frac{C_{\ell xy} L^y_f}{\gamma_y} \right) \frac{L^z_\ell \varepsilon_y}{\gamma_y} \|x - x'\| \\
&+ \frac{C^y_f}{\gamma_y} \left(L^z_{\ell xy} + \frac{C_{\ell xy} L^z_{\ell yy}}{\gamma_y} \right) \varepsilon_y \|x - x'\| + \left(L^z_f \varepsilon_x + \frac{C_{\ell xy} L^z_f \varepsilon_x}{\gamma_y} \right) \frac{C_{\ell xy} + L^z_\ell \varepsilon_y}{\gamma_y} \|x - x'\| \qquad (84) \\
\leq &\left(L_Z^{\varepsilon_x, \varepsilon_y} + \left(L^z_f + \frac{C_{\ell xy} L^z_f}{\gamma_y} \right) \frac{C_{\ell xy} \varepsilon_x}{\gamma_y} \right) \|x - x'\|, \qquad (85)
\end{aligned}
$$

which completes the proof. □

B. Repeated Risk Minimization (Proof of Theorem 1)

Proof. First, let

$$g(\varphi) \triangleq \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(\varphi,x))}[f(\varphi,y^*(\varphi,x);Z)],\ \text{s.t. } y^*(\varphi,x)=\arg\min_{y}\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x)}[\ell(\varphi,y;Z)], \tag{86a}$$

$$g'(\varphi) \triangleq \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(\varphi,x'))}[f(\varphi,y^*(\varphi,x');Z)],\ \text{s.t. } y^*(\varphi,x')=\arg\min_{y}\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x')}[\ell(\varphi,y;Z)]. \tag{86b}$$

Note that $g(\varphi)$ is γ_x-strongly convex and

$$x_{r+1}=R(x_r)\triangleq\arg\min_{\varphi}\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(\varphi,x_r))}[f(\varphi,y^*(\varphi,x_r);Z)],\ \text{s.t. } y^*(\varphi,x_r)=\arg\min_{y}\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x_r)}[\ell(\varphi,y;Z)]. \tag{87}$$

Based on the strong convexity of function $g(\varphi)$, we have

$$g(R(x))-g(R(x'))\geq\langle R(x)-R(x'),\nabla g(R(x'))\rangle+\frac{\gamma_x}{2}\|R(x)-R(x')\|^2, \tag{88a}$$

$$g(R(x'))-g(R(x))\geq\frac{\gamma_x}{2}\|R(x)-R(x')\|^2, \tag{88b}$$

which together give

$$-\gamma_x\|R(x)-R(x')\|^2\geq\langle R(x)-R(x'),\nabla g(R(x'))\rangle. \tag{89}$$

Note that

$$\begin{aligned}
&\|\nabla g(R(x'))-\nabla g'(R(x'))\|\\
\leq&\left\|\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(R(x'),x))}\nabla_x f(R(x'),y^*(R(x'),x);Z)-\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(R(x'),x'))}\nabla_x f(R(x'),y^*(R(x'),x');Z)\right\|\\
&+\left\|\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(R(x'),x)}\nabla^2_{xy}\ell(R(x'),y^*(R(x'),x);Z)\left[\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(R(x'),x)}\nabla^2_{yy}\ell(R(x'),y^*(R(x'),x);Z)\right]^{-1}\right.\\
&\qquad\times\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(R(x'),x))}\nabla_y f(R(x'),y^*(R(x'),x);Z)\\
&\qquad-\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(R(x'),x')}\nabla^2_{xy}\ell(R(x'),y^*(R(x'),x');Z)\left[\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(R(x'),x')}\nabla^2_{yy}\ell(R(x'),y^*(R(x'),x');Z)\right]^{-1}\\
&\qquad\left.\times\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(R(x'),x'))}\nabla_y f(R(x'),y^*(R(x'),x');Z)\right\|
\end{aligned} \tag{90}$$

$$\overset{(73)}{\leq}L_Z^{\varepsilon_x,\varepsilon_y}\|x-x'\|. \tag{91}$$

Using the dual formulation of the optimal transport distance and ε_x-sensitivity of $\mathcal{D}_x$ yields

$$\langle R(x)-R(x'),\nabla g(R(x'))\rangle-\langle R(x)-R(x'),\nabla g'(R(x'))\rangle\geq-L_Z^{\varepsilon_x,\varepsilon_y}\|x-x'\|\|R(x)-R(x')\|. \tag{92}$$

Note that $\langle R(x)-R(x'),\nabla g'(R(x'))\rangle\geq 0$ due to the optimality condition of the UL problem. Combining (89) and (92) yields

$$\|R(x)-R(x')\|\leq\frac{L_Z^{\varepsilon_x,\varepsilon_y}}{\gamma_x}\|x-x'\|\overset{(74)}{=}\left(C_{xy}\varepsilon_x\varepsilon_y+C_y\varepsilon_y\right)\|x-x'\| \tag{93}$$

where

$$C_{xy}\triangleq\frac{L_\ell^z}{\gamma_x\gamma_y}\left(L_f^z+\frac{C_{\ell xy}L_f^z}{\gamma_y}\right), \tag{94a}$$

$$C_y\triangleq\frac{1}{\gamma_x\gamma_y}\left(C_f^y\left(L^z_{\ell xy}+\frac{L^z_{\ell yy}C_{\ell xy}}{\gamma_y}\right)+L_\ell^z\left(\bar{L}_f^x+\frac{C_f^yL^y_{\ell xy}}{\gamma_y}+\frac{C_{\ell xy}}{\gamma_y}\left(L_f^y+\frac{C_f^yL^y_{\ell yy}}{\gamma_y}\right)\right)\right). \tag{94b}$$

By using the definition of the PS solution (3) (i.e., $R(x_{\mathrm{S}})=x_{\mathrm{S}}$) and Bi-RRM (5) (i.e., $x_{r+1}=R(x_r)$), we can conclude that

$$\|x_r-x_{\mathrm{S}}\|_2\leq\left(C_{xy}\varepsilon_x\varepsilon_y+C_y\varepsilon_y\right)\|x_{r-1}-x_{\mathrm{S}}\|\leq\left(C_{xy}\varepsilon_x\varepsilon_y+C_y\varepsilon_y\right)^r\|x_0-x_{\mathrm{S}}\|_2. \tag{95}$$

□

C. Convergence Rate of Bi-SGD (Proof of Theorem 2)

In this section, we will provide detailed proofs of the convergence rate of PP based Bi-SGD.

First, we can have the following descent lemma that quantifies the decrease of the objective value after performing one step update of solving the LL problem.

C.1. Descent Lemma of the LL Problem

Lemma 5. *(Convergence of the LL Variables) Suppose that Assumption 1–Assumption 6 hold. If the iterates* $\{x_r, y_r, \forall r\}$ *are generated by Bi-SGD and the step size* β_r *satisfies*

$$\beta_r \le \frac{\gamma_y}{2(L_\ell^y)^2} \tag{96}$$

then, we have

$$\mathbb{E}\|y_{r+1} - y^*(x_r)\|^2 \le (1-\beta_r\gamma_y)\mathbb{E}\|y_r - y^*(x_r)\|^2 + 2\beta_r^2\sigma_\ell^2, \tag{97}$$

and

$$\begin{aligned}
&\mathbb{E}\|y_{r+1} - y^*(x_{r+1})\|^2\\
&\le \left(1 + Y_1\alpha_r + L_y^{\varepsilon_y}(\sigma'_f+1)\sigma'_f\alpha_r^2\right)\mathbb{E}\|y_{r+1} - y^*(x_r)\|^2 + \left(\frac{C_{\ell xy}L_f^{\varepsilon_x}\alpha_r}{\gamma_y} + Y_2\alpha_r^2\right)\mathbb{E}\|y_r - y^*(x_r)\|^2\\
&\quad + \left(F\alpha_r^2 + \frac{\alpha_r}{2(L_F+\gamma_x)}\right)\left\|\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x_r, y^*(x_r, x^*); Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x^*, y^*(x^*); Z)\right\|^2\\
&\quad + \left(\frac{L_F\gamma_x\alpha_r}{2(L_F+\gamma_x)} + X\alpha_r^2\right)\mathbb{E}\|x_r - x^*\|^2 + 4\left(L_y^{\varepsilon_y}\right)^2\alpha_r^2\sigma_f^2 + L_y^{\varepsilon_y}\alpha_r^2\sigma_f^2 + 4\left(L_y^{\varepsilon_y}\right)^2\alpha_r^2\delta_r^2 + \frac{3}{2}L_y^{\varepsilon_y}\alpha_r\delta_r^2
\end{aligned} \tag{98}$$

where constants Y_1, Y_2, X, F, σ'_f *are as follows*

$$Y_1 \triangleq \frac{C_{\ell xy}L_f^{\varepsilon_x}}{\gamma_y} + \frac{2C_{\ell xy}^2\left(L_Z^{\varepsilon_x,\varepsilon_y}\right)^2(L_F+\gamma_x)}{\gamma_y^2 L_F\gamma_x} + \frac{2C_{\ell xy}^2(L_F+\gamma_x)}{\gamma_y^2} + L_y^{\varepsilon_y}, \tag{99a}$$

$$Y_2 \triangleq 6\left(L_y^{\varepsilon_y}\right)^2(L_f^{\varepsilon_x})^2 + 3(L_f^{\varepsilon_x})^2L_y^{\varepsilon_y}\sigma'_f, \tag{99b}$$

$$X \triangleq 6\left(L_y^{\varepsilon_y}\right)^2\left(L_Z^{\varepsilon_x,\varepsilon_y}\right)^2 + 3L_y^{\varepsilon_y}\left(L_Z^{\varepsilon_x,\varepsilon_y}\right)^2\sigma'_f, \tag{99c}$$

$$F \triangleq 6\left(L_y^{\varepsilon_y}\right)^2 + 3L_y^{\varepsilon_y}\sigma'_f, \quad \textit{and} \quad \sigma'_f \triangleq \sigma_f + \delta_r. \tag{99d}$$

Proof. The distance between y_{r+1} and $y^*(x_r)$ can be quantified as follows.

$$\begin{aligned}
&\mathbb{E}\|y_{r+1} - y^*(x_r)\|^2\\
&= \mathbb{E}\|y_{r+1} - y_r + y_r - y^*(x_r)\|^2 && (100)\\
&\overset{(12a)}{\le} \mathbb{E}\|y_r - y^*(x_r)\|^2 - 2\beta_r\langle y_r - y^*(x_r), \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x_r)}\widehat{\nabla}_y\ell(x_r, y_r; Z)\rangle + \mathbb{E}\|y_{r+1} - y_r\|^2 && (101)\\
&\overset{(a)}{\le} (1-2\beta_r\gamma_y)\mathbb{E}\|y_r - y^*(x_r)\|^2 + \mathbb{E}\|y_{r+1} - y_r\|^2 && (102)\\
&\overset{(12a)}{\le} (1-2\beta_r\gamma_y)\mathbb{E}\|y_r - y^*(x_r)\|^2 + \beta_r^2\mathbb{E}\left\|\widehat{\nabla}\ell(x_r, y_r; Z_y)\right\|^2 && (103)\\
&\overset{(b)}{\le} (1-2\beta_r\gamma_y + 2\beta_r^2(L_\ell^y)^2)\mathbb{E}\|y_r - y^*(x_r)\|^2 + 2\beta_r^2\sigma_\ell^2 && (104)\\
&\overset{(c)}{\le} (1-\beta_r\gamma_y)\mathbb{E}\|y_r - y^*(x_r)\|^2 + 2\beta_r^2\sigma_\ell^2 && (105)
\end{aligned}$$

where (a) is true because $\nabla_y\ell(x_r, y^*(x_r)) = 0$, unbiasedness of the LL gradient estimator, and the strong convexity, i.e.,

$$\langle y_r - y^*(x_r), \nabla_y\ell(x_r, y_r) - \nabla_y\ell(x_r, y^*(x_r))\rangle \ge \gamma_y\|y_r - y^*(x_r)\|^2, \tag{106}$$

and in (b) we use the bounded variance of the LL gradient estimate (i.e., Assumption 6), optimality condition of the LL problem (i.e., $\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x_r)} \nabla_y \ell(x_r, y^*(x_r); Z) = 0$), and Lipschitz continuity, i.e.,

$$\left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_y(x_r)} \nabla_y \ell(x_r, y_r; Z) - \nabla_y \ell(x_r, y^*(x_r); Z) \right\|^2 \leq (L_\ell^y)^2 \|y_r - y^*(x_r)\|^2, \tag{107}$$

and (c) holds when $\beta_r \leq \gamma_y/(2(L_\ell^y)^2)$.

Then, we can decompose $\|y_{r+1} - y^*(x_{r+1})\|^2$ into the following three parts.

$$\begin{aligned}
&\mathbb{E}\|y_{r+1} - y^*(x_{r+1})\|^2 \\
&= \|y_{r+1} - y^*(x_r) + y^*(x_r) - y^*(x_{r+1})\|^2 \qquad (108)\\
&= \mathbb{E}\|y_{r+1} - y^*(x_r)\|^2 + \underbrace{\mathbb{E}\|y^*(x_r) - y^*(x_{r+1})\|^2}_{\triangleq T_1} + 2\underbrace{\mathbb{E}\langle y_{r+1} - y^*(x_r), y^*(x_r) - y^*(x_{r+1})\rangle}_{\triangleq T_2}. \qquad (109)
\end{aligned}$$

Next, we will respectively give the upper bounds of terms T_1 and T_2. Firstly, we have that

$$\begin{aligned}
T_1 &= \mathbb{E}\|y^*(x_r) - y^*(x_{r+1})\|^2 \qquad (110)\\
&\overset{(a)}{\leq} \left(L_y^{\varepsilon_y}\right)^2 \mathbb{E}\|x_{r+1} - x_r\|^2 \qquad (111)\\
&\overset{\text{(12b)}}{\leq} \left(L_y^{\varepsilon_y}\right)^2 \alpha_r^2 \mathbb{E}\|\widehat{\nabla} f(x_r, y_r; Z_x)\|^2 \qquad (112)\\
&\overset{(b)}{\leq} 2\left(L_y^{\varepsilon_y}\right)^2 \alpha_r^2 \|\overline{\nabla} f(x_r, y_r)\|^2 + 4\left(L_y^{\varepsilon_y}\right)^2 \alpha_r^2 \sigma_f^2 + 4\left(L_y^{\varepsilon_y}\right)^2 \alpha_r^2 \delta_r^2 \qquad (113)
\end{aligned}$$

where in (a) we apply the Lipschitz continuity of $y^*(x)$, and (b) follows due to the bounded variance of the UL gradient estimate from Assumption 6.

Let x^* denote x_{S}. Based on the closed form and Lipschitz continuity of $\overline{\nabla} f(x_r, y_r)$, we can obtain

$$\begin{aligned}
\|\overline{\nabla} f(x_r, y_r)\| &= \|\overline{\nabla} f(x_r, y_r) - \overline{\nabla} f(x_r, y^*(x_r)) + \overline{\nabla} f(x_r, y^*(x_r)) - \overline{\nabla} f(x^*, y^*(x^*))\| \qquad (114)\\
&\leq \left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y_r)} \overline{\nabla} f(x_r, y_r; Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x_r))} \overline{\nabla} f(x_r, y^*(x_r); Z) \right\| \\
&\quad + \left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x_r))} \overline{\nabla} f(x_r, y^*(x_r); Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))} \overline{\nabla} f(x_r, y^*(x_r, x^*); Z) \right\| \\
&\quad + \left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))} \overline{\nabla} f(x_r, y^*(x_r, x^*); Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))} \overline{\nabla} f(x^*, y^*(x^*); Z) \right\| \qquad (115)\\
&\leq L_f^{\varepsilon_x} \|y_r - y^*(x_r)\| + L_Z^{\prime\varepsilon_x,\varepsilon_y} \|x_r - x^*\| \\
&\quad + \left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))} \overline{\nabla} f(x_r, y^*(x_r, x^*); Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))} \overline{\nabla} f(x^*, y^*(x^*); Z) \right\| \qquad (116)
\end{aligned}$$

where we use (34) and (73) in the last inequality.

Consequently, we can obtain

$$\begin{aligned}
T_1 \overset{(116)}{\leq} & 2\left(L_y^{\varepsilon_y}\right)^2 \alpha_r^2 \left(3(L_f^{\varepsilon_x})^2 \mathbb{E}\|y_r - y^*(x_r)\|^2 + 3L_Z^{\prime 2} \mathbb{E}\|x_r - x^*\|^2\right) \\
& + 2\left(L_y^{\varepsilon_y}\right)^2 \alpha_r^2 3 \left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))} \nabla f(x_r, y^*(x_r, x^*)) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))} \nabla f(x^*, y^*(x^*)) \right\|^2 \\
& + 4\left(L_y^{\varepsilon_y}\right)^2 \alpha_r^2 \sigma_f^2 + 4\left(L_y^{\varepsilon_y}\right)^2 \alpha_r^2 \delta_r^2. \qquad (117)
\end{aligned}$$

Secondly, we will establish an upper bound for T_2 as follows. Let $\widehat{x}_{r+1} \triangleq \vartheta x_r + (1-\vartheta) x_{r+1}, \vartheta \in [0, 1]$. By the mean-value theorem, we can obtain

$$
\begin{aligned}
T_2 &= \mathbb{E}\langle y^*(x_r) - y_{r+1}, y^*(x_{r+1}) - y^*(x_r)\rangle && (118)\\
&= \mathbb{E}\langle y^*(x_r) - y_{r+1}, \nabla y^*(\widehat{x}_{r+1})^T(x_{r+1} - x_r)\rangle && (119)\\
&= \mathbb{E}\langle y^*(x_r) - y_{r+1}, \alpha_r \nabla y^*(\widehat{x}_{r+1})^T \widehat{\nabla} f(x_r, y_r; Z_x)\rangle && (120)\\
&= T_{21} + T_{22}
\end{aligned}
$$

where

$$
\begin{aligned}
T_{21} &\triangleq \mathbb{E}\langle y^*(x_r) - y_{r+1}, \alpha_r \nabla y^*(\widehat{x}_{r+1})^T \overline{\nabla} f(x_r, y_r; Z_x)\rangle, && (121a)\\
T_{22} &\triangleq \mathbb{E}\left\langle y^*(x_r) - y_{r+1}, \alpha_r \nabla y^*(\widehat{x}_{r+1})^T \left(\widehat{\nabla} f(x_r, y_r; Z_x) - \overline{\nabla} f(x_r, y_r; Z_x)\right)\right\rangle. && (121b)
\end{aligned}
$$

Next, we will provide the upper bounds of T_{11} and T_{22}.

$$
\begin{aligned}
T_{21} &\overset{(54)}{\leq} \frac{C_{\ell xy}}{\gamma_y}\alpha_r \mathbb{E}\|y^*(x_r) - y_{r+1}\|\|\overline{\nabla} f(x_r, y_r)\| && (122)\\
&\overset{(116)}{\leq} \frac{C_{\ell xy}}{\gamma_y}\alpha_r \mathbb{E}\|y^*(x_r) - y_{r+1}\|\Bigg(L_f^{\varepsilon_x}\|y_r - y^*(x_r)\| + L_Z'^{\varepsilon_x,\varepsilon_y}\|x_r - x^*\|\\
&\quad + \left\|\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x_r, y^*(x_r, x^*); Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x^*, y^*(x^*); Z)\right\|\Bigg) && (123)\\
&\leq \frac{C_{\ell xy}}{\gamma_y}L_f^{\varepsilon_x}\alpha_r \mathbb{E}\|y^*(x_r) - y_{r+1}\|\|y_r - y^*(x_r)\| + \frac{C_{\ell xy}}{\gamma_y}L_Z'^{\varepsilon_x,\varepsilon_y}\alpha_r \mathbb{E}\|y^*(x_r) - y_{r+1}\|\|x_r - x^*\|\\
&\quad + \frac{C_{\ell xy}}{\gamma_y}\alpha_r \mathbb{E}\|y^*(x_r) - y_{r+1}\|\left\|\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x_r, y^*(x_r, x^*); Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x^*, y^*(x^*); Z)\right\| && (124)\\
&\overset{(a)}{\leq} \frac{C_{\ell xy}L_f^{\varepsilon_x}\alpha_r}{2\gamma_y}\mathbb{E}\|y_{r+1} - y^*(x_r)\|^2 + \frac{C_{\ell xy}L_f^{\varepsilon_x}\alpha_r}{2\gamma_y}\mathbb{E}\|y_r - y^*(x_r)\|^2\\
&\quad + \frac{C_{\ell xy}^2(L_Z'^{\varepsilon_x,\varepsilon_y})^2(L_F + \gamma_x)\alpha_r}{\gamma_x\gamma_y^2 L_F}\mathbb{E}\|y_{r+1} - y^*(x_r)\|^2 + \frac{L_F\gamma_x\alpha_r}{4(L_F + \gamma_x)}\mathbb{E}\|x_r - x^*\|^2\\
&\quad + \frac{C_{\ell xy}^2(L_F + \gamma_x)\alpha_r}{\gamma_y^2}\mathbb{E}\|y_{r+1} - y^*(x_r)\|^2\\
&\quad + \frac{\alpha_r}{4(L_F + \gamma_x)}\left\|\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x_r, y^*(x_r, x^*); Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x^*, y^*(x^*); Z)\right\|^2 && (125)
\end{aligned}
$$

where in (a) we apply Young's inequality.

Regarding term T_{22}, it can be upper bounded by

$$
\begin{aligned}
T_{22} &= \mathbb{E}\left\langle y^*(x_r) - y_{r+1}, \alpha_r\left(\nabla y^*(\widehat{x}_{r+1}) - \nabla y^*(x_r)\right)^T\left(\widehat{\nabla} f(x_r, y_r; Z_x) - \overline{\nabla} f(x_r, y_r)\right)\right\rangle\\
&\quad + \mathbb{E}\left\langle y^*(x_r) - y_{r+1}, \alpha_r \nabla y^*(x_r)^T\left(\widehat{\nabla} f(x_r, y_r; Z_x) - \overline{\nabla} f(x_r, y_r)\right)\right\rangle && (126)\\
&\overset{(a)}{\leq} L_y^{\varepsilon_y}\alpha_r \mathbb{E}\|y^*(x_r) - y_{r+1}\|\,\|\widehat{x}_{r+1} - x_r\|\left\|\widehat{\nabla} f(x_r, y_r; Z_x) - \overline{\nabla} f(x_r, y_r)\right\|\\
&\quad + L_y^{\varepsilon_y}\alpha_r \mathbb{E}\|y^*(x_r) - y_{r+1}\|\left\|\mathbb{E}\widehat{\nabla} f(x_r, y_r; Z_x) - \overline{\nabla} f(x_r, y_r)\right\| && (127)\\
&\overset{(b)}{\leq} L_y^{\varepsilon_y}\alpha_r \mathbb{E}\|y^*(x_r) - y_{r+1}\|\|x_{r+1} - x_r\|\left\|\widehat{\nabla} f(x_r, y_r; Z_x) - \overline{\nabla} f(x_r, y_r)\right\|\\
&\quad + \frac{L_y^{\varepsilon_y}\alpha_r}{2}\mathbb{E}\|y^*(x_r) - y_{r+1}\|^2 + \frac{L_y^{\varepsilon_y}\alpha_r\delta_r^2}{2} && (128)
\end{aligned}
$$

where (a) holds because of (45) and conditional independence of data sampling of Z_x, in (b) we use (14).

Using the update rule of sequence x_r (12b) in (128) further gives

$$
\begin{aligned}
T_{22} \overset{\text{(12b)}}{\le}\, & L_y^{\varepsilon_y}\alpha_r^2\mathbb{E}\|y^*(x_r)-y_{r+1}\|\left\|\widehat{\nabla} f(x_r,y_r;Z_x)\right\|\left\|\widehat{\nabla} f(x_r,y_r;Z_x)-\overline{\nabla} f(x_r,y_r)\right\| \\
&+\frac{L_y^{\varepsilon_y}\alpha_r}{2}\mathbb{E}\|y^*(x_r)-y_{r+1}\|^2+\frac{L_y^{\varepsilon_y}\alpha_r\delta_r^2}{2} \qquad (129)\\
\overset{(a)}{\le}\, & L_y^{\varepsilon_y}\alpha_r^2\mathbb{E}\|y^*(x_r)-y_{r+1}\|\left\|\overline{\nabla} f(x_r,y_r)\right\|\left\|\widehat{\nabla} f(x_r,y_r;Z_x)-\overline{\nabla} f(x_r,y_r)\right\| \\
&+L_y^{\varepsilon_y}\alpha_r^2\mathbb{E}\|y^*(x_r)-y_{r+1}\|\left\|\widehat{\nabla} f(x_r,y_r;Z_x)-\overline{\nabla} f(x_r,y_r)\right\|^2 \\
&+\frac{L_y^{\varepsilon_y}\alpha_r}{2}\mathbb{E}\|y^*(x_r)-y_{r+1}\|^2+\frac{L_y^{\varepsilon_y}\alpha_r\delta_r^2}{2} \qquad (130)\\
\overset{(b)}{\le}\, & \left(\frac{L_y^{\varepsilon_y}\alpha_r^2(\sigma_f+\delta_r)}{2}+\frac{L_y^{\varepsilon_y}\alpha_r^2(\sigma_f+\delta_r)^2}{2}+\frac{L_y^{\varepsilon_y}\alpha_r}{2}\right)\mathbb{E}\|y^*(x_r)-y_{r+1}\|^2 \\
&+\frac{L_y^{\varepsilon_y}\alpha_r^2(\sigma_f+\delta_r)}{2}\|\overline{\nabla} f(x_r,y_r)\|^2+L_y^{\varepsilon_y}\alpha_r^2\sigma_f^2+\frac{3L_y^{\varepsilon_y}\alpha_r\delta_r^2}{2} \qquad (131)\\
\overset{(116)}{\le}\, & \frac{3L_Z'^2L_y^{\varepsilon_y}\alpha_r^2(\sigma_f+\delta_r)}{2}\mathbb{E}\|x_r-x^*\|^2 \\
&+\frac{3(L_f^{\varepsilon_x})^2L_y^{\varepsilon_y}\alpha_r^2(\sigma_f+\delta_r)}{2}\mathbb{E}\|y_r-y^*(x_r)\|^2 \\
&+\frac{3L_y^{\varepsilon_y}\alpha_r^2(\sigma_f+\delta_r)}{2}\left\|\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x_r,y^*(x_r,x^*);Z)-\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x^*,y^*(x^*);Z)\right\|^2 \\
&+\left(\frac{L_y^{\varepsilon_y}\alpha_r^2(\sigma_f+\delta_r)}{2}+\frac{L_y^{\varepsilon_y}\alpha_r^2(\sigma_f+\delta_r)^2}{2}+\frac{L_y^{\varepsilon_y}\alpha_r}{2}\right)\mathbb{E}\|y^*(x_r)-y_{r+1}\|^2 \\
&+L_y^{\varepsilon_y}\alpha_r^2\sigma_f^2+\frac{3L_y^{\varepsilon_y}\alpha_r\delta_r^2}{2}. \qquad (132)
\end{aligned}
$$

where (a) is true due to the fact that $\|\widehat{\nabla} f(x_r,y_r;Z_x)\|\le\|\overline{\nabla} f(x_r,y_r)\|+\|\widehat{\nabla} f(x_r,y_r;Z_x)-\overline{\nabla} f(x_r,y_r)\|$, and in (b) we use Young's inequality and Assumption 6 (i.e., $\mathbb{E}\|\widehat{\nabla} f(x_r,y_r;Z_x)-\overline{\nabla} f(x_r,y_r)\|\le\sigma_f+\delta_r$).
Combining (125) and (132) yields

$$
\begin{aligned}
T_2 \le\, & \Bigg(\left(\frac{C_{\ell xy}L_f^{\varepsilon_x}}{2\gamma_y}+\frac{C_{\ell xy}^2L_Z'^2(L_F+\gamma_x)}{\gamma_y^2L_F\gamma_x}+\frac{C_{\ell xy}^2(L_F+\gamma_x)}{\gamma_y^2}\right)\alpha_r \\
&+\frac{L_y^{\varepsilon_y}\alpha_r^2(\sigma_f+\delta_r)}{2}+\frac{L_y^{\varepsilon_y}\alpha_r^2(\sigma_f+\delta_r)^2}{2}+\frac{L_y^{\varepsilon_y}\alpha_r}{2}\Bigg)\mathbb{E}\|y_{r+1}-y^*(x_r)\|^2 \\
&+\left(\frac{C_{\ell xy}L_f^{\varepsilon_x}\alpha_r}{2\gamma_y}+\frac{3(L_f^{\varepsilon_x})^2L_y^{\varepsilon_y}\alpha_r^2(\sigma_f+\delta_r)}{2}\right)\mathbb{E}\|y_r-y^*(x_r)\|^2 \\
&+\left(\frac{\alpha_rL_F\gamma_x}{4(L_F+\gamma_x)}+\frac{3L_Z'^2L_y^{\varepsilon_y}\alpha_r^2(\sigma_f+\delta_r)}{2}\right)\mathbb{E}\|x_r-x^*\|^2 \\
&+\left(\frac{3L_y^{\varepsilon_y}\alpha_r^2(\sigma_f+\delta_r)}{2}+\frac{\alpha_r}{4(L_F+\gamma_x)}\right)\left\|\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x_r,y^*(x_r,x^*);Z)-\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x^*,y^*(x^*);Z)\right\|^2 \\
&+L_y^{\varepsilon_y}\alpha_r^2\sigma_f^2+\frac{3L_y^{\varepsilon_y}\alpha_r\delta_r^2}{2}. \qquad (133)
\end{aligned}
$$

To simplify the notation, let $\sigma'_f \triangleq \sigma_f + \delta_r$. Substituting (117) and (133) back into (109) gives

$$\begin{aligned}
&\mathbb{E}\|y_{r+1} - y^*(x_{r+1})\|^2 \\
\leq&\Bigg(1 + \left(\frac{C_{\ell xy} L_f^{\varepsilon_x}}{\gamma_y} + \frac{2C_{\ell xy}^2 L_Z'^2 (L_F + \gamma_x)}{\gamma_y^2 L_F \gamma_x} + \frac{2C_{\ell xy}^2 (L_F + \gamma_x)}{\gamma_y^2}\right)\alpha_r \\
&\quad + L_y^{\varepsilon_y}\alpha_r^2\sigma'_f + L_y^{\varepsilon_y}\alpha_r^2\sigma_f'^2 + L_y^{\varepsilon_y}\alpha_r\Bigg)\mathbb{E}\|y_{r+1} - y^*(x_r)\|^2 \\
&+ \left(6\left(L_y^{\varepsilon_y}\right)^2 (L_f^{\varepsilon_x})^2\alpha_r^2 + \frac{C_{\ell xy} L_f^{\varepsilon_x}\alpha_r}{\gamma_y} + 3(L_f^{\varepsilon_x})^2 L_y^{\varepsilon_y}\alpha_r^2\sigma'_f\right)\mathbb{E}\|y_r - y^*(x_r)\|^2 \\
&+ \left(6\left(L_y^{\varepsilon_y}\right)^2\left(L_Z'^{\varepsilon_x,\varepsilon_y}\right)^2\alpha_r^2 + \frac{\alpha_r L_F\gamma_x}{2(L_F + \gamma_x)} + 3\left(L_Z'^{\varepsilon_x,\varepsilon_y}\right)^2 L_y^{\varepsilon_y}\alpha_r^2\sigma'_f\right)\mathbb{E}\|x_r - x^*\|^2 \\
&+ \left(6\left(L_y^{\varepsilon_y}\right)^2\alpha_r^2 + 3L_y^{\varepsilon_y}\alpha_r^2\sigma'_f + \frac{\alpha_r}{2(L_F + \gamma_x)}\right)\left\|\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x_r, y^*(x_r, x^*); Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x^*, y^*(x^*); Z)\right\|^2 \\
&+ 4\left(L_y^{\varepsilon_y}\right)^2\alpha_r^2\sigma_f^2 + L_y^{\varepsilon_y}\alpha_r^2\sigma_f^2 + 4\left(L_y^{\varepsilon_y}\right)^2\alpha_r^2\delta_r^2 + \frac{3}{2}L_y^{\varepsilon_y}\alpha_r\delta_r^2,
\end{aligned} \tag{134}$$

which gives (98) directly. □

C.2. Descent Lemma of the UL Problem

Lemma 6. *Suppose that Assumption 1–Assumption 6 hold. If the iterates $\{x_r, y_r, \forall r\}$ are generated by Bi-SGD and the ε_x and ε_y satisfy*

$$\boxed{C_x\varepsilon_x + C_{xy}\varepsilon_x\varepsilon_y + C_y\varepsilon_y \le \frac{L_F^{\varepsilon_x,\varepsilon_y}}{4(L_F^{\varepsilon_x,\varepsilon_y}+\gamma_x)} \qquad (135)}$$

where

$$C_x \triangleq \left(L_f^z + \frac{C_{\ell xy}L_f^z}{\gamma_y}\right)\frac{C_{\ell xy}}{\gamma_x\gamma_y}, \tag{136}$$

then, we have

$$\begin{aligned}
&\mathbb{E}\|x_{r+1}-x^*\|^2 \\
\le &\left(1 - \frac{\alpha_r L_F\gamma_x}{L_F+\gamma_x} + \frac{1}{8}\left(\frac{L_F\gamma_x}{L_F+\gamma_x}\alpha_r\right)^2\right)\mathbb{E}\|x_r - x^*\|^2 \\
&+ \left(\frac{4(L_f^{\varepsilon_x})^2(L_F+\gamma_x)\alpha_r}{L_F\gamma_x} + 2(L_f^{\varepsilon_x})^2\alpha_r^2\right)\mathbb{E}\|y_r - y^*(x_r)\|^2 \\
&- \left(\frac{\alpha_r}{L_F+\gamma_x} - 2\alpha_r^2\right)\left\|\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x_r, y^*(x_r,x^*);Z) - \overline{\nabla} f(x^*, y^*(x^*);Z)\right\|^2 \\
&+ 4\alpha_r^2\sigma_f^2 + 4\alpha_r^2\delta_r^2 + \frac{4(L_F+\gamma_x)}{L_F\gamma_x}\alpha_r\delta_r^2.
\end{aligned} \tag{137}$$

Proof. Let x^* denote the x_{S} of this problem. First, we can obtain

$$\begin{aligned}
&\mathbb{E}\|x_{r+1}-x^*\|^2 \\
&= \mathbb{E}\|x_{r+1} - x_r + x_r - x^*\|^2 && (138)\\
&= \mathbb{E}\|x_r - x^*\|^2 + 2\langle x_{r+1}-x_r, x_r - x^*\rangle + \mathbb{E}\|x_{r+1}-x_r\|^2 && (139)\\
&\overset{(12b)}{=} \mathbb{E}\|x_r - x^*\|^2 - 2\alpha_r\left\langle \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y_r)}\widehat{\nabla} f(x_r,y_r;Z), x_r - x^*\right\rangle + \mathbb{E}\|x_{r+1}-x_r\|^2 && (140)\\
&= \mathbb{E}\|x_r - x^*\|^2 - 2\alpha_r\left\langle \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y_r)}\widehat{\nabla} f(x_r,y_r;Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x_r))}\overline{\nabla} f(x_r,y^*(x_r);Z), x_r - x^*\right\rangle \\
&\quad - 2\alpha_r\left\langle \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x_r))}\overline{\nabla} f(x_r,y^*(x_r);Z), x_r - x^*\right\rangle + \mathbb{E}\|x_{r+1}-x_r\|^2. && (141)
\end{aligned}$$

Using the fact that $\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x^*, y^*(x^*);Z) = 0$, we can have

$$\begin{aligned}
&\left\langle \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x_r))}\overline{\nabla} f(x_r,y^*(x_r);Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x_r,y^*(x_r,x^*);Z), x_r - x^*\right\rangle \\
&+ \left\langle \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x_r,y^*(x_r,x^*);Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x^*,y^*(x^*);Z), x_r - x^*\right\rangle \\
&\overset{(a)}{\ge} - L_Z^{\prime\varepsilon_x,\varepsilon_y}\|x_r - x^*\|^2 + \left\langle \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x_r,y^*(x_r,x^*);Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x^*,y^*(x^*);Z), x_r - x^*\right\rangle \\
&\overset{(b)}{\ge} - L_Z^{\prime\varepsilon_x,\varepsilon_y}\mathbb{E}\|x_r - x^*\|^2 + \frac{L_F\gamma_x}{L_F+\gamma_x}\|x_r - x^*\|^2 \\
&+ \frac{1}{L_F+\gamma_x}\left\|\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x_r,y^*(x_r,x^*);Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla} f(x^*,y^*(x^*);Z)\right\|^2,
\end{aligned} \tag{142}$$

where (a) holds by applying Young's inequality and (73), in (b) we use the L_F-smooth and γ_x-strongly convexity of $f(x, y^*(x))$ by applying Lemma 3.11 in (Bubeck et al., 2015).

By substituting (142) into (141), we obtain

$$\begin{aligned}
&\mathbb{E}\|x_{r+1} - x^*\|^2 \\
&\leq \left(1 - \frac{2\alpha_r L_F \gamma_x}{L_F + \gamma_x} + 2\alpha_r L_Z^{\prime\varepsilon_x,\varepsilon_y}\right) \mathbb{E}\|x_r - x^*\|^2 \\
&\quad - 2\alpha_r \left\langle \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y_r)} \widehat{\nabla} f(x_r, y_r; Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x_r))} \overline{\nabla} f(x_r, y^*(x_r); Z), x_r - x^* \right\rangle \\
&\quad + \mathbb{E}\|x_{r+1} - x_r\|^2 - \frac{\alpha_r}{L_F + \gamma_x} \left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))} \overline{\nabla} f(x_r, y^*(x_r, x^*); Z) - \overline{\nabla} f(x^*, y^*(x^*); Z) \right\|^2 \qquad (143) \\
&\overset{(a)}{\leq} \left(1 - \frac{3\alpha_r L_F \gamma_x}{2(L_F + \gamma_x)} + 2\alpha_r L_Z^{\prime\varepsilon_x,\varepsilon_y}\right) \mathbb{E}\|x_r - x^*\|^2 + \frac{4(L_f^{\varepsilon_x})^2 (L_F + \gamma_x)\alpha_r}{L_F \gamma_x} \mathbb{E}\|y_r - y^*(x_r)\|^2 \\
&\quad + \mathbb{E}\|x_{r+1} - x_r\|^2 - \frac{\alpha_r}{L_F + \gamma_x} \left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))} \overline{\nabla} f(x_r, y^*(x_r, x^*); Z) - \overline{\nabla} f(x^*, y^*(x^*); Z) \right\|^2 + \frac{4(L_F + \gamma_x)}{L_F \gamma_x} \alpha_r \delta_r^2 \\
&\qquad (144) \\
&\overset{(b)}{\leq} \left(1 - \frac{3\alpha_r L_F \gamma_x}{2(L_F + \gamma_x)} + 2\alpha_r L_Z^{\prime\varepsilon_x,\varepsilon_y}\right) \mathbb{E}\|x_r - x^*\|^2 + \frac{4(L_f^{\varepsilon_x})^2 (L_F + \gamma_x)\alpha_r}{L_F \gamma_x} \mathbb{E}\|y_r - y^*(x_r)\|^2 \\
&\quad + 2\alpha_r^2 \mathbb{E}\|\overline{\nabla} f(x_r, y_r)\|^2 - \frac{\alpha_r}{L_F + \gamma_x} \left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))} \overline{\nabla} f(x_r, y^*(x_r, x^*)) - \overline{\nabla} f(x^*, y^*(x^*)) \right\|^2 \\
&\quad + 4\alpha_r^2 \sigma_f^2 + 4\alpha_r^2 \delta_r^2 + \frac{4(L_F + \gamma_x)}{L_F \gamma_x} \alpha_r \delta_r^2 \qquad (145) \\
&\overset{(c)}{\leq} \left(1 - \frac{3\alpha_r L_F \gamma_x}{2(L_F + \gamma_x)} + 2\alpha_r L_Z^{\prime\varepsilon_x,\varepsilon_y} + 2(L_Z^{\prime\varepsilon_x,\varepsilon_y} \alpha_r)^2\right) \mathbb{E}\|x_r - x^*\|^2 \\
&\quad + \left(\frac{4(L_f^{\varepsilon_x})^2 (L_F + \gamma_x)\alpha_r}{L_F \gamma_x} + 2(L_f^{\varepsilon_x})^2 \alpha_r^2 \right) \mathbb{E}\|y_r - y^*(x_r)\|^2 \\
&\quad - \left(\frac{\alpha_r}{L_F + \gamma_x} - 2\alpha_r^2 \right) \left\| \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))} \overline{\nabla} f(x_r, y^*(x_r, x^*); Z) - \overline{\nabla} f(x^*, y^*(x^*); Z) \right\|^2 \\
&\quad + 4\alpha_r^2 \sigma_f^2 + 4\alpha_r^2 \delta_r^2 + \frac{4(L_F + \gamma_x)}{L_F \gamma_x} \alpha_r \delta_r^2 \qquad (146)
\end{aligned}$$

where in (a) we use

$$\begin{aligned}
&\left\langle \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y_r)} \overline{\nabla} f(x_r, y_r; Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x_r))} \overline{\nabla} f(x_r, y^*(x_r); Z), x_r - x^* \right\rangle \\
&\quad + \left\langle \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y_r)} \widehat{\nabla} f(x_r, y_r; Z) - \mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y_r)} \overline{\nabla} f(x_r, y_r; Z), x_r - x^* \right\rangle \\
&\leq \frac{L_F \gamma_x}{8(L_F + \gamma_x)} \mathbb{E}\|x_r - x^*\|^2 + \frac{2(L_f^{\varepsilon_x})^2 (L_F + \gamma_x)}{L_F \gamma_x} \mathbb{E}\|y_r - y^*(x_r)\|^2 \\
&\quad + \frac{L_F \gamma_x}{8(L_F + \gamma_x)} \mathbb{E}\|x_r - x^*\|^2 + \frac{2(L_F + \gamma_x)}{L_F \gamma_x} \delta_r^2 \qquad (147) \\
&\leq \frac{L_F \gamma_x}{4(L_F + \gamma_x)} \mathbb{E}\|x_r - x^*\|^2 + \frac{2(L_f^{\varepsilon_x})^2 (L_F + \gamma_x)}{L_F \gamma_x} \mathbb{E}\|y_r - y^*(x_r)\|^2 + \frac{2(L_F + \gamma_x)}{L_F \gamma_x} \delta_r^2, \qquad (148)
\end{aligned}$$

(b) is true due to (12b) and the fact that $\mathbb{E}\|\widehat{\nabla} f(x_r, y_r, Z_x)\| \leq \mathbb{E}\|\widehat{\nabla} f(x_r, y_r, Z_x) - \overline{\nabla} f(x_r, y_r)\| + \|\overline{\nabla} f(x_r, y_r)\| \leq \sigma_f + \delta_r + \|\overline{\nabla} f(x_r, y_r)\|$, and (c) holds by applying (116).

When

$$\frac{L_F\gamma_x}{L_F+\gamma_x} \geq 4L_Z'^{\varepsilon_x,\varepsilon_y}, \tag{149}$$

we have

$$-\alpha_r\frac{L_F\gamma_x}{L_F+\gamma_x} + 2\alpha_r L_Z'^{\varepsilon_x,\varepsilon_y} + 2(L_Z'^{\varepsilon_x,\varepsilon_y}\alpha_r)^2 \leq -\frac{\alpha_r L_F\gamma_x}{2(L_F+\gamma_x)} + \frac{1}{8}\left(\frac{L_F\gamma_x}{L_F+\gamma_x}\alpha_r\right)^2. \tag{150}$$

Note that the expression of $L_Z'^{\varepsilon_x,\varepsilon_y}$ is given in (74). So, the condition

$$\frac{4(L_F+\gamma_x)}{L_F\gamma_x}L_Z'^{\varepsilon_x,\varepsilon_y} \leq 1 \tag{151}$$

is equivalent to

$$\begin{aligned} 1 &\geq \frac{4(L_F+\gamma_x)}{L_F\gamma_x}\left(L_Z^{\varepsilon_x,\varepsilon_y} + \left(L_f^z + \frac{C_{\ell xy}L_f^z}{\gamma_y}\right)\frac{C_{\ell xy}\varepsilon_x}{\gamma_y}\right) && (152)\\ &= \frac{4(L_F+\gamma_x)}{L_F}\left(\frac{L_Z^{\varepsilon_x,\varepsilon_y}}{\gamma_x} + \left(L_f^z + \frac{C_{\ell xy}L_f^z}{\gamma_y}\right)\frac{C_{\ell xy}\varepsilon_x}{\gamma_x\gamma_y}\right) && (153)\\ &\overset{(94)}{=} \frac{4(L_F+\gamma_x)}{L_F}\left(C_x\varepsilon_x + C_{xy}\varepsilon_x\varepsilon_y + C_y\varepsilon_y\right) && (154) \end{aligned}$$

where

$$C_x \triangleq \left(L_f^z + \frac{C_{\ell xy}L_f^z}{\gamma_y}\right)\frac{C_{\ell xy}}{\gamma_x\gamma_y}. \tag{155}$$

Combining (150) and (154) gives the desired result.

□

C.3. Convergence of the Whole Sequence

Proof. Combining (97), (98) and (137), we can get

$$\begin{aligned} &\mathbb{E}\|x_{r+1}-x^*\|^2 + \mathbb{E}\|y_{r+1}-y^*(x_{r+1})\|^2 - \left(\underbrace{\mathbb{E}\|x_r-x^*\|^2 + \mathbb{E}\|y_r-y^*(x_r)\|^2}_{\triangleq\mathcal{P}_r}\right)\\ &\leq \left(-\frac{\alpha_r L_F\gamma_x}{L_F+\gamma_x} + \frac{1}{8}\left(\frac{L_F\gamma_x}{L_F+\gamma_x}\alpha_r\right)^2 + X\alpha_r^2\right)\mathbb{E}\|x_r-x^*\|^2\\ &\quad + \Bigg(\left(1+Y_1\alpha_r + L_y^{\varepsilon_y}(\sigma_f'+1)\sigma_f'\alpha_r^2\right)(1-\beta_r\gamma_y) - 1\\ &\qquad + \left(\frac{C_{\ell xy}L_f^{\varepsilon_x}\alpha_r}{\gamma_y} + Y_2\alpha_r^2 + \frac{4(L_f^{\varepsilon_x})^2(L_F+\gamma_x)\alpha_r}{L_F\gamma_x} + 2(L_f^{\varepsilon_x})^2\alpha_r^2\right)\Bigg)\mathbb{E}\|y_r-y^*(x_r)\|^2\\ &\quad - \left(\frac{\alpha_r}{2(L_F+\gamma_x)} - (2+F)\alpha_r^2\right)\left\|\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla}f(x_r,y^*(x_r,x^*)) - \overline{\nabla}f(x^*,y^*(x^*))\right\|^2\\ &\quad + 4\left(L_y^{\varepsilon_y}\right)^2\alpha_r^2\sigma_f^2 + L_y^{\varepsilon_y}\alpha_r^2\sigma_f^2 + 4\left(L_y^{\varepsilon_y}\right)^2\alpha_r^2\delta_r^2 + \frac{3}{2}L_y^{\varepsilon_y}\alpha_r\delta_r^2 + 4\alpha_r^2\sigma_f^2 + 4\alpha_r^2\delta_r^2 + \frac{4(L_F+\gamma_x)}{L_F\gamma_x}\alpha_r\delta_r^2\\ &\quad + 2\left(1+Y_1\alpha_r + L_y^{\varepsilon_y}(\sigma_f'+1)\sigma_f'\alpha_r^2\right)(1-\beta_r\gamma_y)\beta_r^2\sigma_\ell^2 \end{aligned} \tag{156}$$

where $\mathcal{P}_r$ is the Lyapunov function.

In order to the contraction property of $\mathcal{P}_r$, we need the terms in front of $\mathbb{E}\|x_r-x^*\|^2$, $\mathbb{E}\|y_r-y^*(x_r)\|^2$, $\left\|\mathop{\mathbb{E}}_{Z\sim\mathcal{D}_x(y^*(x^*))}\overline{\nabla}f(x_r,y^*(x_r,x^*);Z) - \overline{\nabla}f(x^*,y^*(x^*);Z)\right\|^2$ to be negative. To be more specific, we have the following requirements.

1) to make the term $-\frac{\alpha_r L_F \gamma_x}{L_F+\gamma_x} + \frac{1}{8}\left(\frac{L_F\gamma_x}{L_F+\gamma_x}\alpha_r\right)^2 + X\alpha_r^2$ be negative, we can choose a small enough step size, i.e.,

$$\begin{aligned} &-\frac{L_F\gamma_x}{2(L_F+\gamma_x)} + \frac{1}{8}\left(\frac{L_F\gamma_x}{L_F+\gamma_x}\right)^2\alpha_r + X\alpha_r \le 0 \\ \implies \alpha_r \le &\frac{L_F\gamma_x}{2(L_F+\gamma_x)\left(\frac{1}{8}\left(\frac{L_F\gamma_x}{L_F+\gamma_x}\right)^2 + X\right)} \le \frac{4L_F\gamma_x}{(L_F+\gamma_x)\left(\frac{L_F\gamma_x}{L_F+\gamma_x}\right)^2} = \frac{4(L_F+\gamma_x)}{L_F\gamma_x} \end{aligned} \tag{157}$$

so that this term is less than $-\frac{L_F\gamma_x}{2(L_F+\gamma_x)}$.

2) to make the term $(1+Y_1\alpha_r+L_y^{\varepsilon_y}(\sigma'_f+1)\sigma'_f\alpha_r^2)(1-\beta_r\gamma_y)-1+(\frac{C_{\ell xy}L_f^{\varepsilon_x}\alpha_r}{\gamma_y}+Y_2\alpha_r^2+\frac{4(L_f^{\varepsilon_x})^2(L_F+\gamma_x)\alpha_r}{L_F\gamma_x}+2(L_f^{\varepsilon_x})^2\alpha_r^2)$ be less than $-\frac{\beta_r\gamma_y}{2}$, we can require

$$\begin{aligned} &Y_1\alpha_r + L_y^{\varepsilon_y}(\sigma'_f+1)\sigma'_f\alpha_r^2 + \left(\frac{C_{\ell xy}L_f^{\varepsilon_x}\alpha_r}{\gamma_y} + Y_2\alpha_r^2 + \frac{4(L_f^{\varepsilon_x})^2(L_F+\gamma_x)\alpha_r}{L_F\gamma_x} + 2(L_f^{\varepsilon_x})^2\alpha_r^2\right) \\ &\le \left(\frac{1}{2} + Y_1\alpha_r + L_y^{\varepsilon_y}(\sigma'_f+1)\sigma'_f\alpha_r^2\right)\beta_r\gamma_y \end{aligned} \tag{158}$$

$$\begin{aligned} \implies \beta_r \ge &\left(\frac{1}{2} + Y_1\alpha_r + L_y^{\varepsilon_y}(\sigma'_f+1)\sigma'_f\alpha_r^2\right)^{-1} \\ &\times \frac{1}{\gamma_y}\left(\left(Y_1 + \frac{C_{\ell xy}L_f^{\varepsilon_x}}{\gamma_y} + \frac{4(L_f^{\varepsilon_x})^2(L_F+\gamma_x)}{L_F\gamma_x}\right)\alpha_r + \left(L_y^{\varepsilon_y}(\sigma'_f+1)\sigma'_f + Y_2 + 2(L_f^{\varepsilon_x})^2\right)\alpha_r^2\right) \end{aligned} \tag{159}$$

$$\ge \frac{1}{\gamma_y}\left(1 + \frac{C_{\ell xy}L_f^{\varepsilon_x}}{\gamma_y Y_1} + \frac{4(L_f^{\varepsilon_x})^2(L_F+\gamma_x)}{L_F\gamma_x Y_1}\right) + Y_1^{-1}\left(L_y^{\varepsilon_y}(\sigma'_f+1)\sigma'_f + Y_2 + 2(L_f^{\varepsilon_x})^2\right)\alpha_r = \Theta(\alpha_r). \tag{160}$$

where the last equality is true because from Lemma 3.2 in (Ghadimi & Wang, 2018) we know that only $\Omega(\log(\alpha_r^{-1}))$ number of data samples can make $\delta_r^2 < \mathcal{O}(\alpha_r)$, so it holds that $\sigma'_f \le \sigma_f + \mathcal{O}(1)$ when we choose $\alpha_r \sim \mathcal{O}(1/r)$.

3) to make the term $\frac{\alpha_r}{2(L_F+\gamma_x)} - (2+F)\alpha_r^2$ be positive, we can require

$$\begin{aligned} \frac{1}{4(L_F+\gamma_x)} - (2+F)\alpha_r \ge 0 \\ \implies \alpha_r \le \frac{1}{4(L_F+\gamma_x)(2+F)}. \end{aligned} \tag{161}$$

Therefore, we can get

$$\mathbb{E}[\mathcal{P}_{r+1}] \le \left(1 - \min\left(\frac{L_F\gamma_x\alpha_r}{2(L_F+\gamma_x)}, \frac{\beta_r\gamma_y}{2}\right)\right)\mathcal{P}_r + \mathcal{O}\left(\alpha_r^2\sigma_f^2\right) + \mathcal{O}\left(\beta_r^2\sigma_\ell^2\right) + \mathcal{O}\left(((\alpha_r+\alpha_r^2)\delta_r^2\right) \tag{162}$$

If we choose $\alpha_r \sim \beta_r$, then

$$\mathbb{E}[\mathcal{P}_{r+1}] \le (1-\Omega(\alpha_r))\,\mathcal{P}_r + \mathcal{O}\left(\alpha_r^2\right) + \mathcal{O}\left((\alpha_r+\alpha_r^2)\delta_r^2\right). \tag{163}$$

Considering that only a number of data samples on the order of $\Omega(\log(\alpha_r^{-1}))$ can result in $\delta_r^2 < \alpha_r$, as stated in Lemma 3.2 of (Ghadimi & Wang, 2018), we can conclude that

$$\mathbb{E}[\mathcal{P}_{r+1}] \le (1-\Omega(\alpha_r))\,\mathcal{P}_r + \mathcal{O}\left(\alpha_r^2\right). \tag{164}$$

If we choose $\alpha_r \sim \beta_r \sim \mathcal{O}(1/r)$, applying the Robbins-Siegmund theorem (Robbins & Siegmund, 1971) gives

$$\lim_{r\to\infty}\|x_r - x_{\mathrm{S}}\| \to 0, \quad \lim_{r\to\infty}\|y_r - y^*(x_{\mathrm{S}})\| \to 0 \quad \textit{almost surely}, \tag{165}$$

since $\sum_{r=1}^{\infty}\alpha_r = \infty$. Also, note that in this choice of the step size $\Pi_{r'=1}^{r}(1-\alpha_{r'}) = \mathcal{O}(1/r)$, so we can have

$$\mathbb{E}[\mathcal{P}_r] = \mathcal{O}\left(\frac{1}{r}\right), \tag{166}$$

which completes the proof. □

C.4. Proof of Corollary 1

Proof. If the full gradient estimate is used, we can immediately from (162) have that

$$\mathcal{P}_{r+1} \leq \left(1 - \min\left(\frac{L_F \gamma_x \alpha_r}{2(L_F + \gamma_x)}, \frac{\beta_r \gamma_y}{2}\right)\right) \mathcal{P}_r, \tag{167}$$

giving a linear convergence of Bi-GD to the BPS solution. □

D. Revisiting BPO and BPS (Proof of Theorem 3)

Proof. Based on the definitions of the BPO solution (2) and BPS solution (3), we know

$$\text{DR}(x_\text{S}, x_\text{S}) \triangleq \underset{Z\sim\mathcal{D}(y^*(x_\text{S}))}{\mathbb{E}} f(x_\text{S}, y^*(x_\text{S}); Z), \text{ s.t. } y^*(x_\text{S}) = \arg\min_y \underset{Z\sim\mathcal{D}(x_\text{S})}{\mathbb{E}} [\ell(x_\text{S}, y; Z)], \tag{168}$$

and

$$\text{DR}(x_\text{O}, x_\text{O}) \triangleq \underset{Z\sim\mathcal{D}(y^*(x_\text{O}))}{\mathbb{E}} f(x_\text{O}, y^*(x_\text{O}); Z), \text{ s.t. } y^*(x_\text{O}) = \arg\min_y \underset{Z\sim\mathcal{D}(x_\text{O})}{\mathbb{E}} [\ell(x_\text{O}, y; Z)]. \tag{169}$$

and

$$\text{DR}(x_\text{O}, x_\text{O}) \leq \text{DR}(x_\text{S}, x_\text{S}) \leq \text{DR}(x_\text{S}, x_\text{O}). \tag{170}$$

Then, we define an auxiliary variable as follows

$$y^*(x_\text{O}, x_\text{S}) = \arg\min_y \underset{Z\sim\mathcal{D}(x_\text{S})}{\mathbb{E}} [\ell(x_\text{O}, y; Z)]. \tag{171}$$

From the strong convexity of $\ell(x, y)$, we know that

$$\left\langle \underset{Z\sim\mathcal{D}(x_\text{S})}{\mathbb{E}} \nabla\ell(x_\text{O}, y^*(x_\text{O})) - \nabla\ell(x_\text{O}, y^*(x_\text{O}, x_\text{S})), y^*(x_\text{O}) - y^*(x_\text{O}, x_\text{S}) \right\rangle \geq \frac{\gamma_y}{2} \|y^*(x_\text{O}) - y^*(x_\text{O}, x_\text{S})\|^2.$$

Similar as the derivations in (49) and (50), according the optimality conditions that both $\underset{Z\sim\mathcal{D}(x_\text{S})}{\mathbb{E}} \nabla\ell(x_\text{O}, y^*(x_\text{O}, x_\text{S}); Z)$ and $\underset{Z\sim\mathcal{D}(x_\text{O})}{\mathbb{E}} \nabla\ell(x_\text{O}, y^*(x_\text{O}); Z)$ are zero, we can obtain

$$\begin{aligned}
&\frac{\gamma_y}{2} \|y^*(x_\text{O}) - y^*(x_\text{O}, x_\text{S})\|^2 \\
&\leq \left\langle \underset{Z\sim\mathcal{D}(x_\text{S})}{\mathbb{E}} \nabla\ell(x_\text{O}, y^*(x_\text{O})) - \underset{Z\sim\mathcal{D}(x_\text{O})}{\mathbb{E}} \nabla\ell(x_\text{O}, y^*(x_\text{O})), y^*(x_\text{O}) - y^*(x_\text{O}, x_\text{S}) \right\rangle &(172)\\
&\leq L_\ell^z \varepsilon_y \|x_\text{S} - x_\text{O}\| \|y^*(x_\text{O}) - y^*(x_\text{O}, x_\text{S})\|, &(173)
\end{aligned}$$

which directly gives

$$\|y^*(x_\text{O}) - y^*(x_\text{O}, x_\text{S})\| \leq \frac{L_\ell^z \varepsilon_y}{\gamma_y} \|x_\text{S} - x_\text{O}\|. \tag{174}$$

Due to the strong convexity of the UL function, we have

$$\begin{aligned}
\underset{Z\sim\mathcal{D}(y^*(x_\text{S}))}{\mathbb{E}} f(x_\text{O}, y^*(x_\text{O}); Z) \geq& \underset{Z\sim\mathcal{D}(y^*(x_\text{S}))}{\mathbb{E}} f(x_\text{S}, y^*(x_\text{S}); Z) \\
&+ \left\langle \underset{Z\sim\mathcal{D}(y^*(x_\text{S}))}{\mathbb{E}} \nabla f(x_\text{S}, y^*(x_\text{S}); Z), x_\text{O} - x_\text{S} \right\rangle + \frac{\gamma_x}{2} \|x_\text{O} - x_\text{S}\|^2. &(175)
\end{aligned}$$

Moverover, by using the optimality condition of the UL variable x_S, we can get

$$\underset{Z\sim\mathcal{D}(y^*(x_\text{S}))}{\mathbb{E}} [f(x_\text{O}, y^*(x_\text{O}); Z) - f(x_\text{S}, y^*(x_\text{S}); Z)] \geq \frac{\gamma_x}{2} \|x_\text{O} - x_\text{S}\|^2. \tag{176}$$

Also, note that

$$\begin{aligned}
&\left| \underset{Z\sim\mathcal{D}(y^*(x_{\mathrm{S}}))}{\mathbb{E}} f(x_{\mathrm{O}}, y^*(x_{\mathrm{O}}); Z) - \underset{Z\sim\mathcal{D}(y^*(x_{\mathrm{O}}))}{\mathbb{E}} f(x_{\mathrm{O}}, y^*(x_{\mathrm{O}}); Z) \right| \\
&\le \varepsilon_x L_z \left\| y^*(x_{\mathrm{S}}) - y^*(x_{\mathrm{O}}, x_{\mathrm{S}}) + y^*(x_{\mathrm{O}}, x_{\mathrm{S}}) - y^*(x_{\mathrm{O}}) \right\| && (177)\\
&\overset{(55),(174)}{\le} \varepsilon_x L_z \left(\frac{C_{\ell xy}}{\gamma_y} \|x_{\mathrm{S}} - x_{\mathrm{O}}\| + \frac{L_\ell^z \varepsilon_y}{\gamma_y} \|x_{\mathrm{S}} - x_{\mathrm{O}}\| \right). && (178)
\end{aligned}$$

By the definition of the PO solution x_{O} shown in (2), we have

$$\underset{Z\sim\mathcal{D}(y^*(x_{\mathrm{O}}))}{\mathbb{E}} f(x_{\mathrm{O}}, y^*(x_{\mathrm{O}}); Z) \le \underset{Z\sim\mathcal{D}(y^*(x_{\mathrm{S}}))}{\mathbb{E}} f(x_{\mathrm{O}}, y^*(x_{\mathrm{S}}); Z), \tag{179}$$

which gives

$$\begin{aligned}
&\underset{Z\sim\mathcal{D}(y^*(x_{\mathrm{S}}))}{\mathbb{E}} \left[f(x_{\mathrm{O}}, y^*(x_{\mathrm{O}}); Z) - f(x_{\mathrm{O}}, y^*(x_{\mathrm{S}}); Z) \right] \\
&\le \underset{Z\sim\mathcal{D}(y^*(x_{\mathrm{S}}))}{\mathbb{E}} f(x_{\mathrm{O}}, y^*(x_{\mathrm{O}}); Z) - \underset{Z\sim\mathcal{D}(y^*(x_{\mathrm{O}}))}{\mathbb{E}} f(x_{\mathrm{O}}, y^*(x_{\mathrm{O}}); Z). && (180)
\end{aligned}$$

Substituting (176) and (178) into (180) yields

$$\frac{\gamma_x}{2} \|x_{\mathrm{O}} - x_{\mathrm{S}}\|^2 \le \frac{\varepsilon_x L_z}{\gamma_y} \left(C_{\ell xy} + L_\ell^z \varepsilon_y \right) \|x_{\mathrm{S}} - x_{\mathrm{O}}\| \tag{181}$$

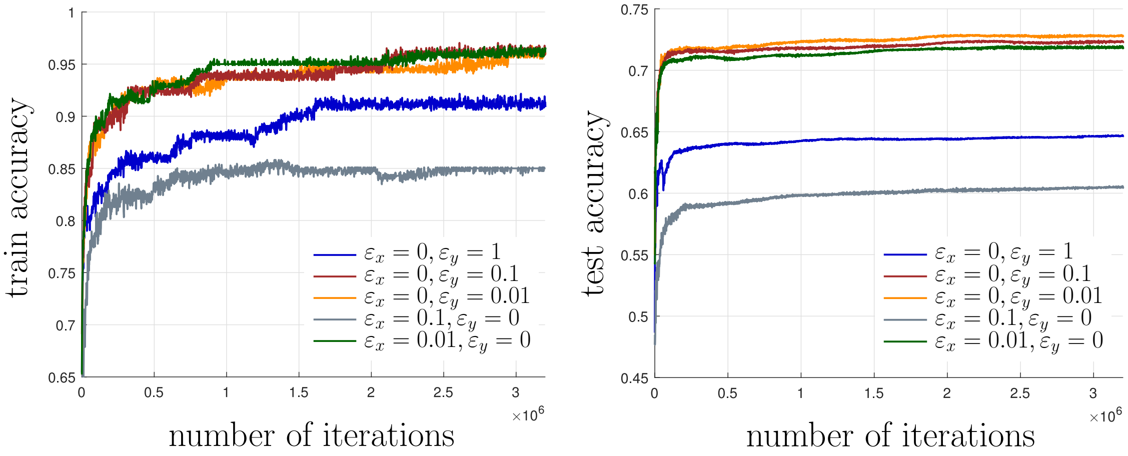
i.e.,

$$\|x_{\mathrm{O}} - x_{\mathrm{S}}\| \le \frac{2\varepsilon_x L_z}{\gamma_x \gamma_y} \left(C_{\ell xy} + L_\ell^z \varepsilon_y \right). \tag{182}$$

□

E. Additional Numerical Experiments

We also evaluate the performance of Bi-SGD for the meta performative prediction learning problem by varying the values of ε_x and ε_y on the spambase data set. As shown in Figure 4, it can be observed that when either the sensitivity parameter ε_x or ε_x is large, Bi-SGD yields lower meta-training and meta-testing accuracies, which aligns with our intuition. Another interesting finding is that the meta-test accuracy with $\varepsilon_x = 0, \varepsilon_y = 0.1$ is significantly higher than the case with $\varepsilon_x = 0.1, \varepsilon_y = 0$, indicating that the meta-testing accuracy is more affected by ε_x compared to ε_y in this example.



(a) Train accuracy

(b) Test accuracy

Figure 4. Comparison of Bi-SGD for meta strategic learning over different combinations of the sensitivity parameters at each level.