Privacy Meets Hierarchy: Differentially Private Distributed Trilevel Learning

Yang Jiao[1] **Kai Yang**[2,3,4*]
[1]School of Computer Science and Engineering, Southeast University, Nanjing, China
[2]School of Computer Science and Technology, Tongji University, Shanghai, China
[3]Shenzhen Loop Area Institute, Shenzhen, China
[4]Shanghai Research Institute for Intelligent Autonomous Systems, Shanghai, China
yang.jiao@seu.edu.cn, kaiyang@tongji.edu.cn

Abstract

Trilevel learning has emerged as a powerful paradigm across various domains in machine learning and networking. In practical scenarios, data is frequently generated and distributed across multiple decentralized nodes. While existing distributed trilevel learning methods circumvent the transmission of raw data, they remain susceptible to privacy leakage. Furthermore, cutting plane methods, which are widely adopted in bilevel and trilevel learning, exhibit analogous intrinsic privacy risks. To address these critical issues, we propose a Differentially Private Distributed Trilevel Optimization (DPDTO) framework. In DPDTO, we first develop a value-function-embedded privacy-preserving cutting plane method for trilevel optimization problems, comprising an inner-layer value function reformulation and an outer-layer cutting plane relaxation. Building upon this, a distributed optimization algorithm is introduced to effectively address trilevel optimization while preserving privacy. Theoretically, we provide a comprehensive analysis for the proposed framework, including asymptotic convergent relaxation, non-asymptotic convergence rate, privacy guarantees, and communication complexity, uncovering a tripartite trade-off among privacy, utility, and complexity in trilevel optimization. Extensive experimental results on two distributed trilevel optimization problems further demonstrate the effectiveness and superiority of the proposed DPDTO.

1 Introduction

Trilevel learning/optimization refers to a class of multilevel optimization problems characterized by a three-tiered hierarchical structure [80]. Recently, it has garnered significant attention in machine learning and networking, encompassing applications such as robust hyperparameter optimization [79], domain adaptation [15], distributionally robust optimization [33], and network defense [109]. The general formulation of a trilevel optimization problem can be expressed as:

$$\begin{aligned} &\min f_1(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \\ &\text{s.t. } \boldsymbol{y} = \arg\min_{\boldsymbol{y}'} f_2(\boldsymbol{x}, \boldsymbol{y}', \boldsymbol{z}) \\ &\quad \text{s.t. } \boldsymbol{z} = \arg\min_{\boldsymbol{z}'} f_3(\boldsymbol{x}, \boldsymbol{y}', \boldsymbol{z}') \\ &\text{var.} \quad \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \end{aligned} \tag{1}$$

where f_1, f_2, f_3 respectively denote the first-, second-, and third-level objective, $\boldsymbol{x} \in \mathbb{R}^{d_1}, \boldsymbol{y} \in \mathbb{R}^{d_2}, \boldsymbol{z} \in \mathbb{R}^{d_3}$ are the variables. Due to the necessity of solving an embedded bilevel optimization subproblem, even *identifying a feasible solution* in trilevel optimization is **NP-hard** [43, 62]. Additionally, in practical applications, data is often generated and distributed across various edge nodes

*Corresponding author: Kai Yang.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

		Convex	Non-convex	Non-linear	Privacy-guaranteed
Jiao et al. [2023]		✓			
Chen et al. [2024]		✓			
Jian et al. [2024]		✓			
Liu et al. [2025]		✓			
Jiao et al. [2024]		✓	✓		
This work		✓	✓	✓	✓

Figure 1: Comparison of the proposed method with state-of-the-art cutting plane methods. Experiments are conducted on the bilevel distributionally robust optimization problem [46], i.e., $\min_{\{p_i,\forall i\}} -\sum_i p_i f_i(w), \text{s.t. } w = \arg\min_w \sum_i p_i f_i(w)$, using the MNIST, Fashion-MNIST, and ECG200 (from the UCR Time Series Archive) datasets.

[34, 106, 111, 117]. To protect data privacy by avoiding the transmission of raw data, distributed trilevel optimization methods have garnered significant attention [42, 44].

However, as demonstrated in previous work, e.g., [105, 77], even without transmitting raw data, there remains a substantial risk of privacy leakage for the private data on each node through the analysis of transmitted information, e.g., gradients [119]. In addition, as the total privacy cost is directly correlated with the number of information queries executed during the optimization process [95, 1], we observe that trilevel optimization is inherently *more vulnerable* to privacy leakage. Specifically, suppose that the lower-level optimization problem is approximated by performing Q steps of gradient descent (this approximation is widely-used in bilevel and trilevel optimization [60, 37, 38]), single-level updates in distributed optimization require only $\mathcal{O}(1)$ gradient queries. In contrast, updating the outermost variable requires $\mathcal{O}(Q)$ queries in bilevel optimization, and *escalates to* $\mathcal{O}(Q^2)$ *in trilevel optimization*. Moreover, providing privacy guarantees for trilevel optimization is important, with motivating applications detailed in Appendix B. Consequently, this raises the first key question of this work: **Can we design an effective differentially private distributed trilevel learning framework?**

Additionally, since cutting plane methods significantly enhance the decomposability of nested problems and provide flexible control over the trade-off between complexity and performance, they are widely adopted in bilevel and trilevel optimization [42, 65, 11, 41, 38]. However, as illustrated in Fig. 1, we observe that existing cutting plane methods are *highly susceptible* to data privacy leakage. This vulnerability stems from the necessity of transmitting local information to update the cutting plane parameters. Furthermore, existing cutting plane methods for trilevel optimization are all based on nested-layer relaxations, which may incur additional approximation errors. This motivates our second key question: **Is it possible to design a privacy-preserving single-layer cutting plane method with theoretical guarantees?**

To this end, we propose DPDTO, a Differentially Private Distributed Trilevel Optimization framework. Specifically, recognizing that the lower-level problem typically serves as a soft constraint for the upper-level objectives, we first leverage the inner-level value function method to reformulate the trilevel optimization problem into a constrained bilevel optimization problem. Then, a privacy-preserving single-layer cutting plane relaxation is introduced. Building upon this, a differentially private distributed algorithm is proposed to efficiently solve trilevel optimization problems while preserving privacy. Theoretically, we demonstrate the asymptotic convergent relaxation, non-asymptotic convergence rate, privacy guarantee, and communication complexity for the proposed DPDTO, and discuss a tripartite trade-off among privacy guarantees, optimization utility, and communication complexity. Importantly, due to the intricate hierarchical dependencies in trilevel optimization and the unavoidable bias introduced by clipping [48, 10, 115], naively integrating differential privacy techniques into existing cutting plane methods or trilevel optimization algorithms is *ineffective* and *fails to* provide any theoretical guarantees. The contributions of this work can be summarized as:

1. In contrast to existing trilevel optimization approaches, a differentially private distributed trilevel optimization framework DPDTO is proposed in this work. To our best knowledge, this is the first work to explore trilevel optimization with differential privacy guarantees.

2. To overcome the privacy vulnerabilities inherent in conventional cutting plane methods, we propose a value-function-embedded privacy-preserving cutting plane method for trilevel optimization. The proposed method not only safeguards data privacy but also effectively mitigates approximation errors.

3. Comprehensive theoretical guarantees of the proposed framework are provided, encompassing convergent relaxation, non-asymptotic convergence rate, privacy preservation, and communication complexity bounds. Empirically, experimental evaluations on two distributed trilevel optimization problems demonstrate the effectiveness and superior performance of the proposed DPDTO.

2 Related Work

2.1 Differential Privacy

Although distributed optimization methods are designed to preserve data privacy by allowing model optimization without sharing local data, recent studies have revealed that they remain vulnerable to privacy breaches caused by indirect leakage [77, 105]. Following [75, 70, 13], a randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ is defined to satisfy (ϵ, δ)-differential privacy if, for all pairs of adjacent datasets $D, D' \in \mathcal{D}$ differing by at most one record ($\|D - D'\|_1 \leq 1$), and for any measurable subset of the output space $S \subseteq \mathcal{R}$, the following probabilistic bound holds:

$$\Pr[\mathcal{M}(D) \in S] \leq \exp(\epsilon) \cdot \Pr[\mathcal{M}(D') \in S] + \delta, \tag{2}$$

where the parameters $\epsilon, \delta > 0$ dictate the privacy level, a smaller ϵ yields stricter privacy, while δ characterizes the probability of the privacy guarantee being violated. To provide privacy guarantees, differential privacy techniques have been widely adopted in the field of distributed optimization. Depending on the network topology, this line of research can be broadly classified into centralized distributed (e.g., [68, 82, 100, 56, 110, 104, 25, 18, 118]) and decentralized differentially private distributed optimization (e.g., [58, 92, 112, 9, 90, 8, 22, 16]). While existing privacy-preserving methods predominantly target single-level problems, the growing prevalence of multilevel optimization [108, 106] necessitates the development of tailored differential privacy frameworks. Recent efforts have begun to bridge this gap, though primarily within the context of bilevel optimization. For instance, to address data privacy in federated meta-learning, the DP-FedMeta architecture was introduced [91]. Similarly, privacy-preserving distributed methods have been developed for non-convex min-max optimization (a special case of bilevel optimization) [77], and decentralized optimization algorithms have been proposed for general bilevel optimization problems in [12]. Nevertheless, these existing approaches are confined to the bilevel paradigm, leaving differentially private distributed trilevel optimization largely **under-explored**. It is worth noting that studying differentially private trilevel optimization is of crucial importance, with *motivating applications* detailed in Appendix B.

2.2 Trilevel Optimization

Trilevel optimization has broad applicability across a diverse range of machine learning domains. For example, [79] formulated a trilevel strategy to concurrently tune hyperparameters and train robust machine learning models. In the realm of late fusion multiview clustering, [55] embedded a min-max paradigm to jointly evaluate perturbation matrices, weight coefficients, and clustering partitions. Besides, TimeAutoAD [40] is proposed to automatically configure the anomaly detection pipeline and optimize the hyperparameters, which can be viewed as a trilevel optimization problem. Beyond these instances, trilevel optimization has found applications in a broad spectrum of problems, including machine translation [31], neural architecture search [2], domain adaptation [15], fairness-aware recommendation [35], image enhancement [26], and out-of-distribution generalization [38]. Since practical applications often involve data scattered across multiple nodes, distributed trilevel optimization has emerged as a crucial strategy to prevent direct data sharing. AFTO [42] established the first algorithmic framework for distributed trilevel optimization by employing a hyper-polyhedral approximation technique. Moreover, a cascaded zeroth-order polynomial approximation strategy was developed in [44], enabling derivative-free solutions for distributed trilevel optimization. However, existing distributed trilevel approaches still face the risk of privacy leakage even without communicating raw data. To mitigate this issue, this work takes the **initial step** in studying distributed trilevel optimization with privacy guarantees and provides comprehensive theoretical guarantees.

3 Method

In practical applications of trilevel optimization, data may be distributed across various workers [54, 72, 89]. In order to avoid direct data transmission among various workers, distributed trilevel optimization has recently attracted increasing research attention, which can be formulated as follows:

$$\begin{aligned} &\min \tfrac{1}{n}\textstyle\sum_{i=1}^{n} f_{1,i}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) \\ &\text{s.t. } \boldsymbol{y}=\arg\min_{\boldsymbol{y}'} \tfrac{1}{n}\textstyle\sum_{i=1}^{n} f_{2,i}(\boldsymbol{x},\boldsymbol{y}',\boldsymbol{z}) \\ &\quad \text{s.t. } \boldsymbol{z}=\arg\min_{\boldsymbol{z}'} \tfrac{1}{n}\textstyle\sum_{i=1}^{n} f_{3,i}(\boldsymbol{x},\boldsymbol{y}',\boldsymbol{z}') \\ &\text{var.} \quad \boldsymbol{x},\boldsymbol{y},\boldsymbol{z}, \end{aligned} \tag{3}$$

where $f_{1,i}, f_{2,i}, f_{3,i}$ respectively denote the first-, second-, and third-level objective in worker i, n represents the number of workers. While existing distributed trilevel methods avoid transmitting raw data, they still suffer from privacy leakage [119, 77]. In this work, we propose the first differentially private trilevel optimization framework. Detailed discussions on the *motivating applications* and *threat model* are provided in Appendix B. For clarity, we outline the proposed framework: First, a value-function-embedded privacy-preserving cutting plane method is proposed in Sec. 3.1, which consists of an inner-level value function reformulation and an outer-level cutting plane relaxation. Then, a differentially private distributed optimization algorithm is introduced in Sec. 3.2 to effectively address the distributed trilevel optimization problems while preserving privacy. Furthermore, comprehensive theoretical guarantees for the proposed framework are provided in Sec. 4.

3.1 Value-Function-Embedded Privacy-Preserving Cutting Plane Relaxation

3.1.1 Inner-Level Value Function Reformulation

In bilevel and trilevel optimization, the lower-level problem is typically viewed as a constraint for the upper-level ones [14, 103, 38]. Accordingly, the third-level optimization problem in (3) can be formulated as the following constraint: $\boldsymbol{z}=\arg\min_{\boldsymbol{z}'} \frac{1}{n}\sum_{i=1}^{n} f_{3,i}(\boldsymbol{x},\boldsymbol{y}',\boldsymbol{z}')$. The value function method has proven highly effective in solving bilevel optimization problems by reformulating the lower-level problem as a value function constraint [60, 99, 107]. Motivated by this, we first consider to equivalently reformulate the third-level optimization problem into the following constraint:

$$\tfrac{1}{n}\textstyle\sum_{i=1}^{n} f_{3,i}(\boldsymbol{x},\boldsymbol{y}',\boldsymbol{z}) - f_{3,i}(\boldsymbol{x},\boldsymbol{y}',\boldsymbol{z}^*) \le 0, \tag{4}$$

where $\boldsymbol{z}^*=\arg\min_{\boldsymbol{z}'} \frac{1}{n}\sum_{i=1}^{n} f_{3,i}(\boldsymbol{x},\boldsymbol{y}',\boldsymbol{z}')$. Nevertheless, there are **two challenges** in constructing value function reformulation for trilevel optimization: (1. *Complexity*) if the exact optimal solution $\boldsymbol{z}^*$ is needed at each iteration, a distributed optimization problem is required to be solved, which incurs significant computational costs. (2. *Privacy*) Even without transmitting raw data, the process of obtaining $\boldsymbol{z}^*$ still faces the risk of privacy leakage, since local information is exchanged when computing $\boldsymbol{z}^*$ (e.g., via distributed SGD [102], ADMM [3], or local SGD [86]). To address these challenges, the proposed approach is motivated by *three factors*: 1) the lower-level problem frequently acts as a *soft constraint* that tolerates a certain degree of relaxation or approximation without resulting in a meaningless solution [46]. 2) it is common to utilize the results after a single step of gradient descent as an approximation for lower-level optimization, e.g., [83, 20, 61, 78, 87]. 3) A single step approach is more *privacy-friendly*, as executing multiple gradient steps would necessitate additional communication rounds [42]. Thus, we propose approximating the optimal solution using a differentially private gradient descent step every K iterations, where $K \ge 1$ is a tunable parameter that provides a flexible trade-off between performance and complexity. Specifically, at t^{th} iteration ($t \bmod K = 0$), worker i computes the differentially private gradients:

$$\tilde{\Delta}_{\boldsymbol{z}}^{i,t} = \tfrac{1}{\tilde{m}}\textstyle\sum_{k=1}^{\tilde{m}} \operatorname{clip}_{c_{\boldsymbol{z}}}(\nabla_{\boldsymbol{z}} f_{3,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t,\xi_{i,k,t})) + \tfrac{2c_{\boldsymbol{z}}}{\tilde{m}}\tilde{v}_{\boldsymbol{z}}^{i,t}, \tag{5}$$

where $\operatorname{clip}_{c_{\boldsymbol{z}}}(\cdot)$ denotes the clipping function with threshold $c_{\boldsymbol{z}}$, $\xi_{i,k,t}$ represents the random variable, $\tilde{m}$ is the batch size, $\tilde{v}_{\boldsymbol{z}}^{i,t} \sim \mathcal{N}(0,\sigma_{\boldsymbol{z}}^2\mathbf{I}_{d_3})$ is the Gaussian noise, and $\frac{2c_{\boldsymbol{z}}}{\tilde{m}}$ is the sensitivity. Subsequently, $\tilde{\Delta}_{\boldsymbol{z}}^{i,t} \in \mathbb{R}^{d_3}$ is sent to the master node. The master then adopts a warm-start strategy [37] to update the value function by computing $\boldsymbol{z}^* = \boldsymbol{z}^t - \eta_{\boldsymbol{z}}/n\sum_{i=1}^{n}\tilde{\Delta}_{\boldsymbol{z}}^{i,t}$, where $\eta_{\boldsymbol{z}}$ is the step size. Consequently, we can obtain a value-function-embedded constrained bilevel optimization problem:

$$\begin{aligned} &\min \tfrac{1}{n}\textstyle\sum_{i=1}^{n} f_{1,i}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) \\ &\text{s.t. } \boldsymbol{y}=\arg\min_{\boldsymbol{y}'} \tfrac{1}{n}\textstyle\sum_{i=1}^{n} f_{2,i}(\boldsymbol{x},\boldsymbol{y}',\boldsymbol{z}) \\ &\quad \text{s.t.} \tfrac{1}{n}\textstyle\sum_{i=1}^{n} f_{3,i}(\boldsymbol{x},\boldsymbol{y}',\boldsymbol{z}) - f_{3,i}(\boldsymbol{x},\boldsymbol{y}',\boldsymbol{z}^*) \le 0 \\ &\text{var.} \quad \boldsymbol{x},\boldsymbol{y},\boldsymbol{z}. \end{aligned} \tag{6}$$

3.1.2 Outer-Level Cutting Plane Relaxation

Cutting plane methods have been widely adopted in bilevel and trilevel optimization owing to their efficacy in augmenting the decomposability of nested structures [65, 44], thereby streamlining the development of distributed optimization algorithms [46, 63]. The core idea behind these methods is to approximate the original feasible region using a set of progressively refined cutting planes. Departing from existing cutting plane approaches, the proposed method makes several **key contributions**: 1) To mitigate the data privacy vulnerabilities inherent in existing approaches, we introduce the first *privacy-preserving* cutting plane method. 2) Whereas existing approaches rely on nested-layer cutting plane relaxation for trilevel optimization, a *single-layer* cutting plane relaxation is employed in the proposed method, effectively reducing approximation errors. 3) In addition, we theoretically demonstrate the proposed method can construct an asymptotic convergent relaxation. Specifically, in problem (6), the lower-level optimization problem can also be regarded as a soft constraint to the upper-level ones. Let functions $h_i(\boldsymbol{x},\boldsymbol{y}',\boldsymbol{z})=f_{3,i}(\boldsymbol{x},\boldsymbol{y}',\boldsymbol{z})-f_{3,i}(\boldsymbol{x},\boldsymbol{y}',\boldsymbol{z}^*),\forall i$, building upon prior works [107, 81] and applying Lagrangian relaxation, the second-level optimization problem can be formulated as $\boldsymbol{y}=\arg\min_{\boldsymbol{y}'}\frac{1}{n}\sum_{i=1}^n g_{2,i}(\boldsymbol{x},\boldsymbol{y}',\boldsymbol{z})$, where $g_{2,i}(\boldsymbol{x},\boldsymbol{y}',\boldsymbol{z})=f_{2,i}(\boldsymbol{x},\boldsymbol{y}',\boldsymbol{z})+\vartheta h_i(\boldsymbol{x},\boldsymbol{y}',\boldsymbol{z})$ and ϑ is the Lagrange multiplier. Furthermore, this problem is equivalent to the constraint $\phi(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})\leq 0$, where $\phi(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})=||\boldsymbol{y}-\arg\min_{\boldsymbol{y}'}\frac{1}{n}\sum_{i=1}^n g_{2,i}(\boldsymbol{x},\boldsymbol{y}',\boldsymbol{z})||^2$. Drawing inspiration from the efficacy of cutting plane methods, we consider to relax the lower-level optimization problem by introducing a set of cutting planes:

$$\boldsymbol{\mathcal{C}}=\left\{\boldsymbol{x}^\top\text{diag}(\boldsymbol{a}_{x,l})\boldsymbol{x}+\boldsymbol{y}^\top\text{diag}(\boldsymbol{a}_{y,l})\boldsymbol{y}+\boldsymbol{z}^\top\text{diag}(\boldsymbol{a}_{z,l})\boldsymbol{z}+\boldsymbol{b}_{x,l}^\top\boldsymbol{x}+\boldsymbol{b}_{y,l}^\top\boldsymbol{y}+\boldsymbol{b}_{z,l}^\top\boldsymbol{z}\leq c_l,\forall l\right\},\tag{7}$$

where $\boldsymbol{a}_{x,l}\in\mathbb{R}^{d_1},\boldsymbol{a}_{y,l}\in\mathbb{R}^{d_2},\boldsymbol{a}_{z,l}\in\mathbb{R}^{d_3},\boldsymbol{b}_{x,l}\in\mathbb{R}^{d_1},\boldsymbol{b}_{y,l}\in\mathbb{R}^{d_2},\boldsymbol{b}_{z,l}\in\mathbb{R}^{d_3}$ and $c_l\in\mathbb{R}^1$ are the parameters in l^{th} cutting plane. Nevertheless, similar to the development of the inner-level value function, the construction of outer-level cutting planes also faces *challenges regarding complexity and privacy*. To address the challenges, we propose a privacy-preserving cutting plane method with theoretical guarantees. The proposed method refines the relaxation by updating the cutting planes every K iterations, a process that involves pruning inactive cutting planes and introducing new ones. Specifically, at the t^{th} iteration where $t \bmod K=0$, worker i first computes:

$$\tilde{\Delta}_{\boldsymbol{y}}^{i,t}=\frac{1}{\tilde{m}}\sum_{k=1}^{\tilde{m}}\text{clip}_{c_{\boldsymbol{y}}}(\nabla_{\boldsymbol{y}}g_{2,i}(\boldsymbol{x}^t,\boldsymbol{y}_0,\boldsymbol{z}^t,\xi_{i,k,t}))+\frac{2c_{\boldsymbol{y}}}{\tilde{m}}\tilde{v}_{\boldsymbol{y}}^{i,t},\tag{8}$$

where $\tilde{v}_{\boldsymbol{y}}^{i,t}\sim\mathcal{N}(0,\sigma_{\boldsymbol{y}}^2\mathbf{I}_{d_2})$ is the Gaussian noise, $\tilde{m}$ is the batch size, $\text{clip}_{c_{\boldsymbol{y}}}(\cdot)$ is the clipping function with threshold $c_{\boldsymbol{y}}$ to bound the sensitivity $\frac{2c_{\boldsymbol{y}}}{\tilde{m}}$. Subsequently, the local updates $\tilde{\Delta}_{\boldsymbol{y}}^{i,t}\in\mathbb{R}^{d_2}$ are transmitted to the master node, which computes $\hat{V}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)=\boldsymbol{y}^t-\boldsymbol{y}_0+\frac{\eta_{\boldsymbol{y}}}{n}\sum_{i=1}^n\tilde{\Delta}_{\boldsymbol{y}}^{i,t}$ and $\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)=||\hat{V}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)||^2$. Next, the term $\bar{\Delta}^t=\frac{2\eta_{\boldsymbol{y}}}{n}\hat{V}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)\in\mathbb{R}^{d_2}$ is broadcast back to the workers. Upon receiving this information from the master, each worker i computes the noisy local Hessian-vector product as follows:

$$\begin{aligned}
HVX_i^t&=\tfrac{1}{\tilde{m}}\textstyle\sum_{k=1}^{\tilde{m}}\nabla_{\boldsymbol{xy}}^2g_{2,i}(\boldsymbol{x}^t,\boldsymbol{y}_0,\boldsymbol{z}^t,\xi_{i,k,t})\cdot\boldsymbol{w}_{i,k}^t+\tfrac{2L\|\bar{\Delta}^t\|}{\tilde{m}}\tilde{v}_{\boldsymbol{x}}^{i,t},\\
HVZ_i^t&=\tfrac{1}{\tilde{m}}\textstyle\sum_{k=1}^{\tilde{m}}\nabla_{\boldsymbol{zy}}^2g_{2,i}(\boldsymbol{x}^t,\boldsymbol{y}_0,\boldsymbol{z}^t,\xi_{i,k,t})\cdot\boldsymbol{w}_{i,k}^t+\tfrac{2L\|\bar{\Delta}^t\|}{\tilde{m}}\tilde{v}_{\boldsymbol{z}}^{i,t},\\
\boldsymbol{w}_{i,k}^t&=\begin{cases}\bar{\Delta}^t,\text{if }||\nabla_{\boldsymbol{y}}g_{2,i,k,t}||\leq c_{\boldsymbol{y}},\\ \frac{c_{\boldsymbol{y}}}{||\nabla_{\boldsymbol{y}}g_{2,i,k,t}||}\left(\bar{\Delta}^t-\frac{\nabla_{\boldsymbol{y}}g_{2,i,k,t}(\nabla_{\boldsymbol{y}}g_{2,i,k,t}^\top\bar{\Delta}^t)}{||\nabla_{\boldsymbol{y}}g_{2,i,k,t}||^2}\right),\text{if }||\nabla_{\boldsymbol{y}}g_{2,i,k,t}||>c_{\boldsymbol{y}},\end{cases}
\end{aligned}\tag{9}$$

where we denote $\nabla_{\boldsymbol{y}}g_{2,i,k,t}=\nabla_{\boldsymbol{y}}g_{2,i}(\boldsymbol{x}^t,\boldsymbol{y}_0,\boldsymbol{z}^t,\xi_{i,k,t})$ for brevity, and $\tilde{v}_{\boldsymbol{x}}^{i,t}\sim\mathcal{N}(0,\sigma_{\boldsymbol{x}}^2\mathbf{I}_{d_1})$. Once computed, the noisy local Hessian-vector products $HVX_i^t\in\mathbb{R}^{d_1}$ and $HVZ_i^t\in\mathbb{R}^{d_3}$ are uploaded to the master node, yielding: $\mathbb{E}[\frac{1}{n}\sum_{i=1}^n HVX_i^t]=\frac{\partial\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)}{\partial\boldsymbol{x}},2\hat{V}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)=\frac{\partial\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)}{\partial\boldsymbol{y}},\mathbb{E}[\frac{1}{n}\sum_{i=1}^n HVZ_i^t]=\frac{\partial\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)}{\partial\boldsymbol{z}}$. Thus, by combining the obtained information in the master, the new differentially private cutting plane $\mathbb{C}_{\text{DP}}^{\text{new}}$ will be generated as follows:

$$\begin{aligned}
&\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)+\begin{bmatrix}\frac{1}{n}\sum_{i=1}^n HVX_i^t\\ 2\hat{V}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)\\ \frac{1}{n}\sum_{i=1}^n HVZ_i^t\end{bmatrix}^\top\begin{bmatrix}\boldsymbol{x}-\boldsymbol{x}^t\\ \boldsymbol{y}-\boldsymbol{y}^t\\ \boldsymbol{z}-\boldsymbol{z}^t\end{bmatrix}-\tfrac{L}{2}||\boldsymbol{x}-\boldsymbol{x}^t||^2\\
&-\tfrac{L}{2}||\boldsymbol{y}-\boldsymbol{y}^t||^2-\tfrac{L}{2}||\boldsymbol{z}-\boldsymbol{z}^t||^2-\eta_{\boldsymbol{y}}^2(2(1+2\vartheta)\sqrt{G_f}D_1+D_1^2)-4\eta_{\boldsymbol{y}}D_1D_{\boldsymbol{y}}\\
&\leq\eta_{\boldsymbol{y}}\left(2D_1+16L\sqrt{D_{\boldsymbol{y}}}\right)D_2+\eta_{\boldsymbol{y}}^2\left(16L(1+2\vartheta)\sqrt{G_f}+12LD_1\right)D_2,
\end{aligned}\tag{10}$$

Algorithm 1 DPDTO: Differentially Private Distributed Trilevel Optimization

Initialization: iteration $t = 0$, variables $\boldsymbol{x}^0$, $\boldsymbol{y}^0$, $\boldsymbol{z}^0$.
repeat
  for *local worker* i **do**
    updates local variables $\boldsymbol{x}^{i,t+1}, \boldsymbol{y}^{i,t+1}, \boldsymbol{z}^{i,t+1}$ in Gauss-Seidel manner based on Eqs. (13)-(15);
    transmits the computed $\Delta_{\boldsymbol{x}}^{i,t}$, $\Delta_{\boldsymbol{y}}^{i,t}$, $\Delta_{\boldsymbol{z}}^{i,t}$ to the master.
  end for
  for *master* **do**
    updates the global variables according to Eq. (16);
    broadcasts the updated global variables $\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}$ to the workers.
  end for
  if $t \mod K == 0$ **then**
    updates inner-level value function according to Eqs. (5)-(6);
    generates new cutting plane $\mathbb{C}_{\mathrm{DP}}^{\mathrm{new}}$ according to Eqs. (8)-(10);
    adds new cutting plane and removes inactive cutting plane according to Eq. (11).
  end if
  $t = t + 1$;
until termination.

where $D_1 = \sqrt{(1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2 \sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}}$, $D_2 = \sqrt{D_{\boldsymbol{x}}} + \sqrt{D_{\boldsymbol{y}}} + \sqrt{D_{\boldsymbol{z}}}$, G_f and L are constants, which will be discussed in Sec. 4, ϑ is the Lagrangian multiplier, and $\eta_{\boldsymbol{y}}$ is the step-size. Subsequently, the set of cutting planes is updated by adding the new $\mathbb{C}_{\mathrm{DP}}^{\mathrm{new}}$ and removing the inactive cutting planes:

$$\boldsymbol{\mathcal{C}}^{t+1} = \boldsymbol{\mathcal{C}}^t \cup \mathbb{C}_{\mathrm{DP}}^{\mathrm{new}}, \quad \boldsymbol{\mathcal{C}}^{t+1} = \begin{cases} \boldsymbol{\mathcal{C}}^{t+1} \setminus \mathbb{C}_{\mathrm{DP}}^{l}, \text{if } o_l(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t) \leq 0, \\ \boldsymbol{\mathcal{C}}^{t+1}, \text{otherwise}, \end{cases} \tag{11}$$

where $o_l(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \boldsymbol{x}^\top \mathrm{diag}(\boldsymbol{a}_{x,l})\boldsymbol{x} + \boldsymbol{y}^\top \mathrm{diag}(\boldsymbol{a}_{y,l})\boldsymbol{y} + \boldsymbol{z}^\top \mathrm{diag}(\boldsymbol{a}_{z,l})\boldsymbol{z} + \boldsymbol{b}_{x,l}^\top \boldsymbol{x} + \boldsymbol{b}_{y,l}^\top \boldsymbol{y} + \boldsymbol{b}_{z,l}^\top \boldsymbol{z} - c_l$ represents the l^{th} cutting plane constraint. Theoretically, we demonstrate in Theorem 4.3 that the proposed privacy-preserving cutting plane method yields a convergent relaxation of the original feasible region, with detailed discussions presented in Remark D.1. Consequently, the resulting value-function-embedded privacy-preserving cutting plane relaxation problem can be formulated as:

$$\begin{aligned} &\min \tfrac{1}{n} \textstyle\sum_{i=1}^n f_{1,i}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \\ &\text{s.t.} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix}^\top \mathrm{diag}\left(\begin{bmatrix} \boldsymbol{a}_{x,l} \\ \boldsymbol{a}_{y,l} \\ \boldsymbol{a}_{z,l} \end{bmatrix} \right) \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix} + \begin{bmatrix} \boldsymbol{b}_{x,l} \\ \boldsymbol{b}_{y,l} \\ \boldsymbol{b}_{z,l} \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix} \leq c_l, \forall l \\ &\text{var.} \quad \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}. \end{aligned} \tag{12}$$

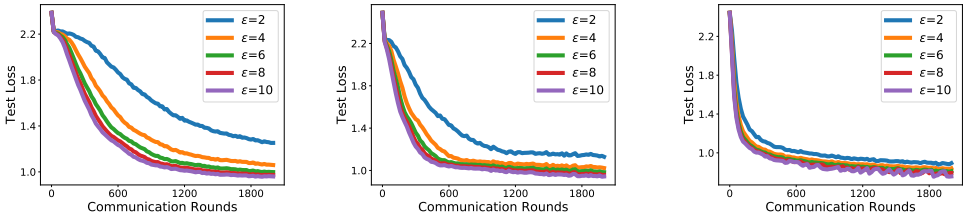
3.2 Differentially Private Distributed Optimization

In this section, a differentially private distributed optimization method is proposed to effectively address the trilevel optimization problem. First, inspired by previous studies on bilevel optimization [107, 67, 81, 7], we apply Lagrangian relaxation to problem (12) to obtain the Lagrangian function $L_\lambda(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \frac{1}{n}\sum_i L_{\lambda,i}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \frac{1}{n}\sum_i f_{1,i}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) + \sum_l \lambda o_l(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, where λ is the Lagrangian multiplier. Due to the use of the cutting plane method, the decomposability of the highly nested trilevel optimization problem is significantly enhanced, thereby enabling it to be solved in a distributed manner. Specifically, in $(t+1)^{\mathrm{th}}$ iteration, at worker i, the local variables for the three levels are updated in a *Gauss-Seidel manner*, dictated by the hierarchical dependencies of the optimization problem. It first computes the gradient with respect to variable $\boldsymbol{x}$ and clips it at threshold $c_{\boldsymbol{x}}$, then the Gaussian noise $v_{\boldsymbol{x}}^{i,t} \sim \mathcal{N}(0, \sigma_{\boldsymbol{x}}^2 \mathbf{I}_{d_1})$ calibrated to the sensitivity $\frac{2c_{\boldsymbol{x}}}{m}$ is added:

$$\begin{aligned} g_{\boldsymbol{x}}^{i,t} &= \tfrac{1}{m}\textstyle\sum_{k=1}^m \mathrm{clip}_{c_{\boldsymbol{x}}}(\nabla_{\boldsymbol{x}} L_{\lambda,i}(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t; \xi_{i,k})), \\ \Delta_{\boldsymbol{x}}^{i,t} &= g_{\boldsymbol{x}}^{i,t} + \tfrac{2c_{\boldsymbol{x}}}{m} v_{\boldsymbol{x}}^{i,t}, \boldsymbol{x}^{i,t+1} = \boldsymbol{x}^t - \eta_{\boldsymbol{x}} \Delta_{\boldsymbol{x}}^{i,t}, \end{aligned} \tag{13}$$

where $\eta_{\boldsymbol{x}}$ is the step-size, $\xi_{i,k}$ is the random variable, m is the batch size. Exploiting the latest state of the first-level variable, the second-level variable $\boldsymbol{y}$ is updated. Conditioned on $\boldsymbol{x}^{i,t+1}$, this step follows an analogous perturbed gradient descent trajectory:

$$\begin{aligned} g_{\boldsymbol{y}}^{i,t} &= \tfrac{1}{m}\textstyle\sum_{k=1}^m \mathrm{clip}_{c_{\boldsymbol{y}}}(\nabla_{\boldsymbol{y}} L_{\lambda,i}(\boldsymbol{x}^{i,t+1}, \boldsymbol{y}^t, \boldsymbol{z}^t; \xi_{i,k})), \\ \Delta_{\boldsymbol{y}}^{i,t} &= g_{\boldsymbol{y}}^{i,t} + \tfrac{2c_{\boldsymbol{y}}}{m} v_{\boldsymbol{y}}^{i,t}, \boldsymbol{y}^{i,t+1} = \boldsymbol{y}^t - \eta_{\boldsymbol{y}} \Delta_{\boldsymbol{y}}^{i,t}, \end{aligned} \tag{14}$$



(a) MNIST (b) QMNIST (c) Fashion MNIST

Figure 2: The impact of various privacy budget ϵ on test loss in various datasets.

Table 1: The comparisons between the proposed DPDTO with the state-of-the-art methods.

Dataset	Setting	DPNFL	DP+ASPIRE	DP+FAT	DP+FBO	LDPDSBO	DP+AFTO	**DPDTO**
MNIST	$\epsilon=4$	18.89±2.45%	29.86±1.22%	57.04±0.84%	56.15±0.69%	58.12±1.04%	61.64±1.26%	**65.91±0.45%**
	$\epsilon=8$	19.72±2.38%	31.05±1.05%	58.93±0.80%	58.79±0.61%	59.70±0.65%	66.31±0.36%	**68.96±0.28%**
F-MNIST	$\epsilon=4$	24.22±1.67%	32.31±2.57%	52.17±0.34%	52.48±0.45%	53.33±0.86%	57.81±0.86%	**62.57±1.38%**
	$\epsilon=8$	26.55±0.82%	36.26±2.45%	53.72±0.56%	53.41±0.49%	54.71±0.85%	61.11±1.01%	**66.36±1.04%**
QMNIST	$\epsilon=4$	19.13±2.23%	29.51±1.20%	56.15±1.11%	55.98±1.07%	56.11±0.63%	64.21±1.70%	**66.93±0.39%**
	$\epsilon=8$	19.77±2.21%	30.55±1.23%	56.96±0.54%	57.17±0.94%	57.56±0.74%	66.57±1.35%	**68.40±0.41%**
USPS	$\epsilon=4$	16.13±1.58%	22.38±2.13%	54.78±0.76%	54.83±0.57%	55.82±0.97%	56.73±0.51%	**60.29±0.60%**
	$\epsilon=8$	20.29±1.73%	32.47±2.36%	55.74±0.34%	56.54±0.59%	56.81±0.64%	60.96±0.46%	**63.93±0.48%**

where $v_{\boldsymbol{y}}^{i,t} \sim \mathcal{N}(0, \sigma_{\boldsymbol{y}}^2 \mathbf{I}_{d_2})$ and $\eta_{\boldsymbol{y}}$ are the Gaussian noise and step-size, respectively. Subsequently, the third-level variable $\boldsymbol{z}$ is updated based on the newly acquired $\boldsymbol{x}^{i,t+1}$ and $\boldsymbol{y}^{i,t+1}$:

$$\begin{aligned} &g_{\boldsymbol{z}}^{i,t} = \frac{1}{m}\sum_{k=1}^{m} \text{clip}_{c_{\boldsymbol{z}}}(\nabla_{\boldsymbol{z}} L_{\lambda,i}(\boldsymbol{x}^{i,t+1}, \boldsymbol{y}^{i,t+1}, \boldsymbol{z}^t; \xi_{i,k})), \\ &\Delta_{\boldsymbol{z}}^{i,t} = g_{\boldsymbol{z}}^{i,t} + \frac{2c_{\boldsymbol{z}}}{m} v_{\boldsymbol{z}}^{i,t}, \boldsymbol{z}^{i,t+1} = \boldsymbol{z}^t - \eta_{\boldsymbol{z}} \Delta_{\boldsymbol{z}}^{i,t}. \end{aligned} \tag{15}$$

Upon the completion of these local computations, the obtained $\Delta_{\boldsymbol{x}}^{i,t} \in \mathbb{R}^{d_1}$, $\Delta_{\boldsymbol{y}}^{i,t} \in \mathbb{R}^{d_2}$, and $\Delta_{\boldsymbol{z}}^{i,t} \in \mathbb{R}^{d_3}$ are transmitted to the master. The master node then performs a global aggregation to update the global model parameters and broadcast back the updated parameters to workers:

$$\begin{aligned} &\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_{\boldsymbol{x}} \bar{g}_{\boldsymbol{x}}^t, \boldsymbol{y}^{t+1} = \boldsymbol{y}^t - \eta_{\boldsymbol{y}} \bar{g}_{\boldsymbol{y}}^t, \boldsymbol{z}^{t+1} = \boldsymbol{z}^t - \eta_{\boldsymbol{z}} \bar{g}_{\boldsymbol{z}}^t, \\ &\bar{g}_{\boldsymbol{x}}^t = \frac{1}{n}\sum_{i=1}^{n} \Delta_{\boldsymbol{x}}^{i,t}, \bar{g}_{\boldsymbol{y}}^t = \frac{1}{n}\sum_{i=1}^{n} \Delta_{\boldsymbol{y}}^{i,t}, \bar{g}_{\boldsymbol{z}}^t = \frac{1}{n}\sum_{i=1}^{n} \Delta_{\boldsymbol{z}}^{i,t}. \end{aligned} \tag{16}$$

The details of the proposed method are summarized in Algorithm 1, and its theoretical guarantees are provided in the next section.

4 Theoretical Analysis

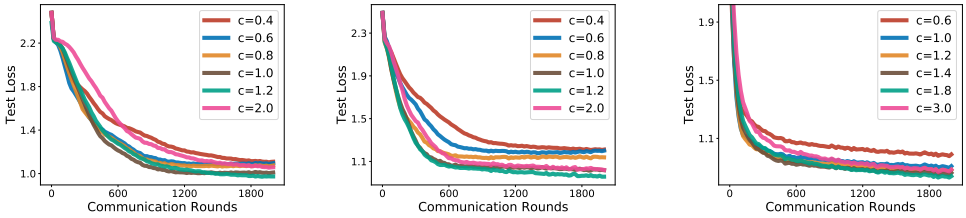
Assumption 4.1 ***(Smoothness)*** *Following the prior work in the related fields, e.g., [93, 120, 30, 49, 64, 45], we assume that the gradient of each level function is L-Lipschitz continuous, that is, for any $\boldsymbol{w}, \boldsymbol{w}'$, we have $||\nabla f_{b,i}(\boldsymbol{w}) - \nabla f_{b,i}(\boldsymbol{w}')|| \leq L||\boldsymbol{w} - \boldsymbol{w}'||, b = 1, 2, 3$.*

Assumption 4.2 ***(Boundedness and Unbiased Estimation)*** *Following previous work in related fields, e.g., [17, 76, 53, 85, 5], the optimization variables are assumed to be bounded in the iterative process, i.e., $||\boldsymbol{x}||^2 \leq D_{\boldsymbol{x}}, ||\boldsymbol{y}||^2 \leq D_{\boldsymbol{y}}, ||\boldsymbol{z}||^2 \leq D_{\boldsymbol{z}}$. Consistent with prior works, e.g., [114, 59, 92, 74, 57], we assume the gradients are bounded, i.e., $\mathbb{E}[||\nabla f_{b,i}(\boldsymbol{w}, \xi)||^2] \leq G_f, b = 1, 2, 3$, and the gradient estimators are assumed to be unbiased, i.e., $\mathbb{E}[\nabla f_{b,i}(\boldsymbol{w}, \xi)] = \nabla f_{b,i}(\boldsymbol{w}), \forall b$, following [13, 75, 32].*

Theorem 4.3 ***(Asymptotic Convergence of the Cumulative Relaxation)*** *Let $\mathcal{C}^t$ denote the cumulative feasible region by retaining all generated privacy-preserving cutting planes at t^{th} iteration. It follows that $\mathcal{C}^t$ is a relaxation of the original feasible region when $\phi(\cdot)$ is L-smooth, and it converges as more cutting planes are added:*

$$\begin{aligned} &\mathcal{H} = \{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} | \phi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \leq 0\} \subseteq \mathcal{C}^t = \{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} | \mathbb{E}[o_l(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})] \leq 0, \forall l\}, \\ &\mathcal{H} \subseteq \lim_{t\to\infty} \downarrow \mathcal{C}^t \subseteq \cdots \subseteq \mathcal{C}^t \subseteq \mathcal{C}^{t-1} \subseteq \cdots \subseteq \mathcal{C}^0. \end{aligned} \tag{17}$$

Proof: *Please refer to Appendix D. More discussions are offered in Remark D.1.*



(a) MNIST (b) QMNIST (c) Fashion MNIST

Figure 3: The impact of various clipping thresholds with the setting $\epsilon = 4$.

Table 2: Trade-off among complexity, privacy, and performance. MSE of noised data is reported.

	Iran		Productivity		Boston		QSAR		Diabetes		Heart	
	ϵ=2.0	ϵ=0.5	ϵ=2.0	ϵ=0.5	ϵ=2.0	ϵ=0.5	ϵ=2.0	ϵ=0.5	ϵ=2.0	ϵ=0.5	ϵ=2.0	ϵ=0.5
$K=1$	0.1045	0.1086	0.0596	0.0613	0.3817	0.3973	0.6254	0.6732	0.5371	0.5815	0.1677	0.1763
$K=5$	0.1066	0.1099	0.0698	0.0796	0.3857	0.4012	0.6309	0.6763	0.5416	0.5853	0.1691	0.1782
$K=10$	0.1148	0.1174	0.0814	0.1094	0.3916	0.4107	0.6418	0.6829	0.5454	0.5930	0.1704	0.1823

Theorem 4.4 ***(Non-Asymptotic Convergence of DPDTO)*** *Under Assumptions 4.1 and 4.2, let step-sizes $\eta_{\boldsymbol{x}} = \eta_{\boldsymbol{y}} = \eta_{\boldsymbol{z}} = \frac{1}{\sqrt{T}}$, we can derive:*

$$\begin{aligned}&\textstyle\sum_{t=0}^{T-1} \frac{||\nabla L_\lambda(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)||^2}{T} \leq \frac{(4L_\lambda(\boldsymbol{x}_0,\boldsymbol{y}_0,\boldsymbol{z}_0)-f^*)}{\sqrt{T}} + \left(\frac{2G_{\boldsymbol{x}}^4}{c_{\boldsymbol{x}}^2} + \frac{4G_{\boldsymbol{y}}^4}{c_{\boldsymbol{y}}^2} + \frac{4G_{\boldsymbol{z}}^4}{c_{\boldsymbol{z}}^2}\right)\\ &+\frac{2\hat{L}^2(c_{\boldsymbol{x}}^2+c_{\boldsymbol{y}}^2+c_{\boldsymbol{z}}^2)}{\sqrt{T}} + \frac{8\hat{L}^2}{\sqrt{T}}\left(\frac{d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{nm^2} + \frac{d_2c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{nm^2} + \frac{d_3c_{\boldsymbol{z}}^2\sigma_{\boldsymbol{z}}^2}{nm^2}\right)\\ &+\frac{\hat{L}^2}{T}\left(18c_{\boldsymbol{x}}^2 + 12c_{\boldsymbol{y}}^2 + \frac{24d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{nm^2} + \frac{16d_2c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{nm^2} + \frac{48d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{m^2} + \frac{32d_2c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{m^2}\right),\end{aligned} \tag{18}$$

where $c_{\boldsymbol{x}}, c_{\boldsymbol{y}}, c_{\boldsymbol{z}}$ are clipping thresholds, d_1, d_2, d_3 are the dimensions of each level variables, $f^* = \min L_\lambda(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, $\hat{L}$ is a constant related to L-smoothness, $G_{\boldsymbol{x}}, G_{\boldsymbol{y}}, G_{\boldsymbol{z}}$ are the constants related to the gradient boundedness G_f, m is the batch size, n is the number of workers in distributed systems.

Proof: *Please refer to Appendix E. And further discussions are provided in Remark E.4.*

Theorem 4.5 ***(Privacy of DPDTO)*** *There exist constants β_0, β_1 for any privacy budget $\epsilon \leq 2\beta_0\beta_1\left(3q_1^2T + 4q_2^2\left\lfloor\frac{T}{K}\right\rfloor\right)$ with sampling rates $q_1 = m/M$ and $q_2 = \tilde{m}/M$, where M represents the total number of data samples [13, 115], the proposed algorithm is (ϵ, δ)-differentially private for any $\delta > 0$ when $\min\{\sigma_{\boldsymbol{x}}, \sigma_{\boldsymbol{y}}, \sigma_{\boldsymbol{z}}\} \geq \sqrt{3q_1^2T + 4q_2^2\left\lfloor\frac{T}{K}\right\rfloor}\frac{2\sqrt{\beta_0\log(1/\delta)}}{\epsilon}$.*

Proof: *Please refer to Appendix F. Furthermore, Corollary F.1 provides a detailed characterization of the **tripartite trade-off** among privacy, utility, and complexity.*

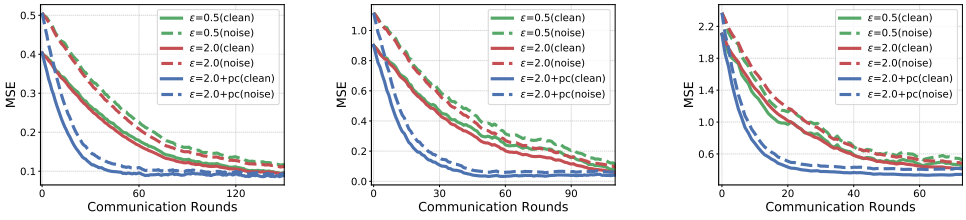
Theorem 4.6 ***(Communication Complexity of DPDTO)*** *The communication complexity is typically measured as the total number of bits exchanged throughout the entire iterative process [84, 50]. Accordingly, the communication complexity of the proposed method can be formulated as:*

$$64n(d_1 + d_2 + d_3)T + 32n(d_1 + 2d_2 + 3d_3)\lfloor T/K\rfloor. \tag{19}$$

Proof: *Please refer to Appendix G.*

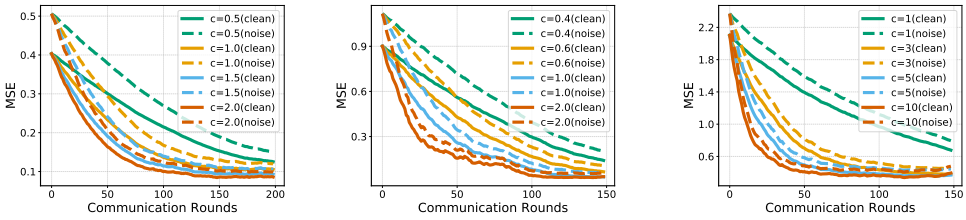
5 Experiment

To evaluate the performance of the proposed DPDTO, experiments are conducted on two distributed trilevel optimization problems: distributed distributionally robust optimization and distributed robust hyperparameter optimization. We compare the proposed DPDTO with various methods, including the differentially private distributed single-level method DPNFL [6], as well as state-of-the-art bilevel and trilevel methods integrated with differential privacy technique, i.e., LDPDSBO [12], DP+ASPIRE [46], DP+FAT [121], DP+FBO [107], and DP+AFTO [42]. Please note that a naive combination of distributed trilevel optimization with differential privacy technique (DP+AFTO) is included solely for empirical comparison, as such an approach lacks theoretical guarantees. Due to space limitation, detailed experimental settings are provided in Appendix A.



(a) Iran (b) Productivity (c) Boston

Figure 4: The impact of different privacy budget and periodic communication on DPDTO.



(a) Iran (b) Productivity (c) Boston

Figure 5: The impact of various clipping thresholds in DRHO problem.

5.1 Distributed Distributionally Robust Optimization (DDRO)

Given the established success of DDRO in tackling data heterogeneity [73, 17, 46], an advanced DDRO framework to further counter adversarial vulnerabilities [96] is introduced, which can be formulated as a distributed trilevel optimization problem. Due to space limitation, the problem formulation and detailed experimental settings are offered in Appendix A.1. In line with previous DDRO works [39, 76], the digit recognition tasks using MNIST, QMNIST, Fashion-MNIST, and USPS datasets are considered. To demonstrate two-level robustness, we follow the evaluation standards of [46, 17] by reporting the worst-case worker's robustness against adversarial attack [69].

Effectiveness. First, we evaluate the proposed DPDTO against the state-of-the-art methods. As shown in Table 1, DPDTO achieves superior performance, demonstrating its effectiveness in solving the trilevel optimization problem while preserving privacy. This improvement stems from two key factors. First, unlike the distributed single-level and bilevel methods, the proposed DPDTO is a trilevel optimization framework inherently capable of addressing highly nested problems. Second, compared to state-of-the-art trilevel optimization method DP+AFTO, the proposed method offers two key advantages: 1) it represents the first differentially private trilevel optimization framework with theoretical guarantees for both convergence and privacy. 2) A theoretically guaranteed privacy-preserving cutting plane method is proposed, which utilizes a single-layer cutting plane relaxation rather than the nested-layer method in DP+AFTO, thereby effectively reducing approximation errors.

Impact of privacy budgets and clipping thresholds. Furthermore, we investigate the impact of various privacy budgets ϵ on the performance of the proposed DPDTO. It can be seen from Fig. 2 that the performance improves when the privacy guarantees are relaxed (i.e., with a larger ϵ), indicating the trade-off between privacy and utility [115]. In addition, we evaluate the performance of DPDTO under various clipping thresholds. As illustrated in Fig. 3, the performance initially improves as the clipping threshold increases within a certain range. This is because an overly small clipping threshold leads to excessive truncation, which introduces additional bias. However, it can also be observed that setting the clipping threshold too large adversely affects the results. Since the sensitivity is directly tied to the clipping threshold, a larger threshold necessitates the injection of more noise to maintain the same level of privacy guarantee, which is consistent with the discussion in [95].

5.2 Distributed Robust Hyperparameter Optimization (DRHO)

DRHO [42, 44, 79] provides a principled approach to jointly tuning hyperparameters while simultaneously training a model resilient to perturbations, which is a distributed trilevel optimization problem.

The detailed formulation, experimental setting, and additional results are reported in Appendix A.2. Consistent with the settings of [79, 42], the MSE is reported for both clean and noised test data across various regression datasets: Diabetes, Boston, Iran, Productivity, QSAR, and Heart.

Impact of privacy budgets and clipping thresholds. We evaluate DPDTO under various privacy budgets ϵ. As illustrated in Figs. 4 and 11, a relatively larger ϵ yields better convergence, and DPDTO remains effective even when a strong privacy guarantee is required, i.e., $\epsilon = 0.5$. Then, we evaluate the impact of clipping thresholds. It is seen from Figs. 5 and 12 that appropriately increasing the clipping threshold enhances convergence rate, since an excessively small threshold introduces additional bias.

Effectiveness and efficiency. Additionally, the results in Figs. 4 and 11 show that DPDTO can be efficiently integrated with periodic communication to further enhance communication efficiency. Besides, we compare DPDTO with the state-of-the-art distributed trilevel learning method DP+AFTO. It is shown in Fig. 10 that DPDTO outperforms DP+AFTO in both effectiveness and efficiency, since DPDTO is the first theoretically guaranteed differentially private trilevel optimization framework, and a privacy-preserving cutting plane method with reduced approximation errors is introduced.

Tripartite Trade-off among complexity, utility, and privacy. In DPDTO, we can flexibly adjust K, a parameter governing the update of cutting planes and value function, to balance the trade-off between optimization utility and complexity. As established in Theorem 4.6, the communication complexity is inversely related to K, meaning a larger K reduces the communication overhead. Conversely, the empirical results in Table 2 reveal that a smaller K yields superior optimization utility, as it provides a better value-function-embedded cutting plane relaxation. Furthermore, we observe that higher utility can be achieved with relaxed privacy guarantees. Achieving a stricter privacy guarantee (i.e., a smaller ϵ) necessitates injecting larger noise, which inevitably degrades the utility unless communication constraints are relaxed (e.g., using a smaller K). This intricate interplay highlights a fundamental *tripartite trade-off* among privacy guarantees, optimization utility, and communication complexity, which is consistent with our analysis in Corollary F.1.

6 Conclusion

In this work, we present DPDTO, the first distributed trilevel optimization framework equipped with privacy guarantees. Within this framework, we develop a value-function-embedded privacy-preserving cutting plane method, which is then integrated with a distributed optimization algorithm to efficiently solve trilevel optimization problems while safeguarding data privacy. Theoretically, we establish the asymptotic convergent relaxation, non-asymptotic convergence rate, privacy guarantee, and communication complexity for the proposed framework. Extensive experimental results not only demonstrate the effectiveness and superiority of the proposed DPDTO, but also illuminate the tripartite trade-off among privacy, optimization utility, and communication complexity.

Acknowledgments and Disclosure of Funding

The work of Yang Jiao was supported by the Start-up Research Fund of Southeast University RF1028626120. The work of Kai Yang was supported in part by NSFC under Grant 12371519 and in part by Shenzhen Loop Area Institute under Project FP202602. The authors declare no competing interests.

References

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] S. Ali and M. A. Wani. Tri-level optimization for gradient-based neural architecture search. In *2024 International Conference on Machine Learning and Applications (ICMLA)*, pages 1546–1552. IEEE, 2024.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[4] D. Byrd and A. Polychroniadou. Differentially private secure multi-party computation for federated learning in financial applications. In *Proceedings of the first ACM international conference on AI in finance*, pages 1–9, 2020.

[5] J. Cao, R. Jiang, N. Abolfazli, E. Yazdandoost Hamedani, and A. Mokhtari. Projection-free methods for stochastic simple bilevel optimization with convex lower-level problem. *Advances in Neural Information Processing Systems*, 36, 2024.

[6] L. Chen, X. Ding, Z. Bao, P. Zhou, and H. Jin. Differentially private federated learning on non-iid data: Convergence analysis and adaptive optimization. *IEEE Transactions on Knowledge and Data Engineering*, 36(9):4567–4581, 2024.

[7] P. Chen, X. Shi, R. Jiang, and J. Wang. Penalty-based methods for simple bilevel optimization under hölderian error bounds. *Advances in Neural Information Processing Systems*, 37: 140731–140765, 2024.

[8] Q. Chen, Z. Wang, M. Yan, H. Yan, X. Lin, and J. Zhou. Defedgcn: privacy-preserving decentralized federated gcn for recommender system. *IEEE Transactions on Services Computing*, 18(2):729–742, 2025.

[9] S. Chen, D. Yu, Y. Zou, J. Yu, and X. Cheng. Decentralized wireless federated learning with differential privacy. *IEEE Transactions on Industrial Informatics*, 18(9):6273–6282, 2022.

[10] X. Chen, S. Z. Wu, and M. Hong. Understanding gradient clipping in private sgd: A geometric perspective. *Advances in neural information processing systems*, 33:13773–13782, 2020.

[11] X. Chen, Y. Xiong, and K. Yang. Robust beamforming for downlink multi-cell systems: A bilevel optimization perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[12] Z. Chen and Y. Wang. Locally differentially private decentralized stochastic bilevel optimization with guaranteed convergence accuracy. In *Forty-first International Conference on Machine Learning*, 2024.

[13] A. Cheng, P. Wang, X. S. Zhang, and J. Cheng. Differentially private federated learning with local regularization and sparsification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10122–10131, 2022.

[14] R. W. T. Chew, Q. P. Nguyen, and B. K. H. Low. Bilbo: Bilevel bayesian optimization. In *International Conference on Machine Learning*, pages 10249–10268. PMLR, 2025.

[15] S. K. Choe, W. Neiswanger, P. Xie, and E. Xing. Betty: An automatic differentiation library for multilevel optimization. In *The Eleventh International Conference on Learning Representations*, 2023.

[16] E. Cyffers, M. Even, A. Bellet, and L. Massoulié. Muffliato: Peer-to-peer privacy amplification for decentralized optimization and averaging. *Advances in Neural Information Processing Systems*, 35:15889–15902, 2022.

[17] Y. Deng, M. M. Kamani, and M. Mahdavi. Distributionally robust federated averaging. *Advances in neural information processing systems*, 33:15111–15122, 2020.

[18] S. Denisov, H. B. McMahan, J. Rush, A. Smith, and A. Guha Thakurta. Improved differential privacy for sgd via optimal private linear operators on adaptive streams. *Advances in Neural Information Processing Systems*, 35:5910–5924, 2022.

[19] A. Fallah, A. Mokhtari, and A. Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in neural information processing systems*, 33:3557–3568, 2020.

[20] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[21] N. Firdaus and Z. Raza. Densefedcnn: a federated learning-based framework with client-specific hyperparameter tuning for multi-class brain mri classification. *International Journal of Machine Learning and Cybernetics*, 17(3):93, 2026.

[22] R. Francis and S. P. Chepuri. Differentially private and communication-efficient decentralized learning using deep quantizers. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[23] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd USENIX Security Symposium*, pages 17–32. USENIX Association, 2014.

[24] A. Ganesh, M. Haghifam, T. Steinke, and A. Guha Thakurta. Faster differentially private convex optimization via second-order methods. *Advances in Neural Information Processing Systems*, 36:79426–79438, 2023.

[25] F. Gao, R. Huang, and J. Yang. Federated online prediction from experts with differential privacy: separations and regret speed-ups. *Advances in Neural Information Processing Systems*, 37:125547–125587, 2024.

[26] J. Gao and Y. Liu. Enhancing images with coupled low-resolution and ultra-dark degradations: A tri-level learning framework. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8642–8651, 2024.

[27] T. Giovannelli, G. D. Kent, and L. N. Vicente. A stochastic gradient method for trilevel optimization. *arXiv preprint arXiv:2505.06805*, 2025.

[28] P. Guo, D. Yang, A. Hatamizadeh, A. Xu, Z. Xu, W. Li, C. Zhao, D. Xu, S. Harmon, E. Turkbey, et al. Auto-fedrl: Federated hyperparameter optimization for multi-institutional medical image segmentation. In *European Conference on Computer Vision*, pages 437–455. Springer, 2022.

[29] A. Hatamizadeh, H. Yin, P. Molchanov, A. Myronenko, W. Li, P. Dogra, A. Feng, M. G. Flores, J. Kautz, D. Xu, et al. Do gradient inversion attacks make federated learning unsafe? *IEEE Transactions on Medical Imaging*, 42(7):2044–2056, 2023.

[30] Y. He, J. Hu, X. Huang, S. Lu, B. Wang, and K. Yuan. Distributed bilevel optimization with communication compression. In *Proceedings of the 41st International Conference on Machine Learning*, pages 17877–17920, 2024.

[31] Y. He, R. Zhang, S. A. Somayajula, and P. Xie. Transformer architecture search for improving out-of-domain generalization in machine translation. *Transactions on Machine Learning Research*, 2024.

[32] R. Hu, Y. Guo, and Y. Gong. Federated learning with sparsified model perturbation: Improving accuracy under client-level differential privacy. *IEEE Transactions on Mobile Computing*, 23 (8):8242–8255, 2023.

[33] L. Huang, J. Zhu, D. Zhang, Y. Chen, and Y. Li. A tri-level distributionally robust optimization model against coordinated cyber-physical attacks considering attack resource uncertainties. *Electrical Engineering*, 107(8):10253–10270, 2025.

[34] M. Huang, D. Zhang, and K. Ji. Achieving linear speedup in non-iid federated bilevel learning. In *International conference on machine learning*, pages 14039–14059. PMLR, 2023.

[35] X. Huang, J. Lian, H. Wang, H. Liao, D. Lian, and X. Xie. A data-centric multi-objective learning framework for responsible recommendation systems. In *Proceedings of the ACM Web Conference 2024*, pages 3129–3138, 2024.

[36] J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

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

[38] C. Jian, K. Yang, and Y. Jiao. Tri-level navigator: Llm-empowered tri-level learning for time series ood generalization. *Advances in Neural Information Processing Systems*, 37: 110613–110642, 2024.

[39] Y. Jiao, K. Yang, and D. Song. Distributed distributionally robust optimization with non-convex objectives. *Advances in neural information processing systems*, 35:7987–7999, 2022.

[40] Y. Jiao, K. Yang, D. Song, and D. Tao. Timeautoad: Autonomous anomaly detection with self-supervised contrastive loss for multivariate time series. *IEEE Transactions on Network Science and Engineering*, 9(3):1604–1619, 2022.

[41] Y. Jiao, K. Yang, T. Wu, D. Song, and C. Jian. Asynchronous distributed bilevel optimization. In *The Eleventh International Conference on Learning Representations*, 2023.

[42] Y. Jiao, K. Yang, T. Wu, C. Jian, and J. Huang. Provably convergent federated trilevel learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12928–12937, 2024.

[43] Y. Jiao, X. Wang, and K. Yang. Pr-attack: Coordinated prompt-rag attacks on retrieval-augmented generation in large language models via bilevel optimization. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 656–667, 2025.

[44] Y. Jiao, K. Yang, and C. Jian. Dtzo: Distributed trilevel zeroth order learning with provable non-asymptotic convergence. In *Forty-second International Conference on Machine Learning*, 2025.

[45] Y. Jiao, K. Jiao, K. Yang, N. Aitsaadi, I. Fajjari, et al. First-order constrained trilevel optimization over distributed networks for robust coreset selection. *arXiv preprint arXiv:2607.27632*, 2026.

[46] Y. Jiao, K. Yang, and D. Song. Federated distributionally robust optimization with non-convex objectives: Algorithm and analysis. *IEEE Transactions on Mobile Computing*, 25(1): 1219–1235, 2026.

[47] Z. Ke, Y. Ming, X.-P. Nguyen, C. Xiong, and S. Joty. Demystifying domain-adaptive post-training for financial llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31021–31047, 2025.

[48] A. Koloskova, H. Hendrikx, and S. U. Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, pages 17343–17363. PMLR, 2023.

[49] B. Kong, S. Zhu, S. Lu, X. Huang, and K. Yuan. Decentralized bilevel optimization: A perspective from transient iteration complexity. *Journal of Machine Learning Research*, 26 (240):1–64, 2025.

[50] J. H. Korhonen and D. Alistarh. Towards tight communication lower bounds for distributed optimisation. *Advances in Neural Information Processing Systems*, 34:7254–7266, 2021.

[51] J. Le, D. Zhang, X. Lei, L. Jiao, K. Zeng, and X. Liao. Privacy-preserving federated learning with malicious clients and honest-but-curious servers. *IEEE Transactions on Information Forensics and Security*, 18:4329–4344, 2023.

[52] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

[53] Y. Lei and K. Tang. Stochastic composite mirror descent: Optimal bounds with high probabilities. *Advances in Neural Information Processing Systems*, 31, 2018.

[54] A. Li, G. Wang, M. Hu, J. Sun, L. Zhang, L. A. Tuan, and H. Yu. Joint client-and-sample selection for federated learning via bi-level optimization. *IEEE Transactions on Mobile Computing*, 2024.

[55] M. Li, X. Liu, Y. Zhang, and W. Liang. Late fusion multiview clustering via min-max optimization. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9417–9427, 2023.

[56] Y. Li, S. Wang, and T.-H. Chang. Differentially private federated stochastic primal-dual learning for internet of vehicles. *IEEE Internet of Things Journal*, 12(11):17034–17050, 2025.

[57] Z. Li, H. Zhao, B. Li, and Y. Chi. Soteriafl: A unified framework for private federated learning with communication compression. *Advances in Neural Information Processing Systems*, 35: 4285–4300, 2022.

[58] W. Lin, B. Li, and C. Wang. Towards private learning on decentralized graphs with local differential privacy. *IEEE Transactions on Information Forensics and Security*, 17:2936–2946, 2022.

[59] X. Ling, J. Fu, K. Wang, H. Liu, and Z. Chen. Ali-dpfl: Differentially private federated learning with adaptive local iterations. In *2024 IEEE 25th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pages 349–358. IEEE, 2024.

[60] B. Liu, M. Ye, S. Wright, P. Stone, and Q. Liu. Bome! bilevel optimization made easy: A simple first-order approach. *Advances in Neural Information Processing Systems*, 35: 17248–17262, 2022.

[61] H. Liu, K. Simonyan, and Y. Yang. Darts: Differentiable architecture search. In *International Conference on Learning Representations*, 2018.

[62] R. Liu, J. Gao, J. Zhang, D. Meng, and Z. Lin. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10045–10067, 2021.

[63] Y. Liu and K. Yang. Asynchronous decentralized federated anomaly detection for 6g networks. *IEEE Transactions on Cognitive Communications and Networking*, 2025.

[64] Y. Liu, K. Yang, Y. Zhu, K. Yang, C. Jian, W. Ni, X. Ye, and Y. Ouyang. A nested zeroth-order fine-tuning approach for cloud-edge llm agents. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 3–15. Springer, 2025.

[65] Y. Liu, K. Yang, Y. Zhu, K. Yang, and H. Zhao. Argus: Federated non-convex bilevel learning over 6 g space-air-ground integrated network. *IEEE Transactions on Network Science and Engineering*, pages 1–17, 2025.

[66] A. Lowy and D. Liu. Differentially private bilevel optimization: Efficient algorithms with near-optimal rates. volume 38, pages 27168–27209, 2026.

[67] Z. Lu and S. Mei. First-order penalty methods for bilevel optimization. *SIAM Journal on Optimization*, 34(2):1937–1969, 2024.

[68] H. Ma, K. Yang, and Y. Jiao. Cellular traffic prediction via byzantine-robust asynchronous federated learning. *IEEE Transactions on Network Science and Engineering*, 2025.

[69] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[70] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2018.

[71] C. Meijer, J. Huang, S. Sharma, E. Lazovik, and L. Y. Chen. Ts-inverse: a gradient inversion attack tailored for federated time series forecasting models. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 110–124. IEEE, 2025.

[72] N. Mitic, A. Pyrgelis, and S. Sav. Privacy-preserving hyperparameter tuning for federated learning. *IEEE Transactions on Privacy*, 2:1–14, 2025.

[73] M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019.

[74] T. Murata and T. Suzuki. Diff2: Differential private optimization via gradient differences for nonconvex distributed learning. In *International Conference on Machine Learning*, pages 25523–25548. PMLR, 2023.

[75] M. Noble, A. Bellet, and A. Dieuleveut. Differentially private federated learning on heterogeneous data. In *International conference on artificial intelligence and statistics*, pages 10110–10145. PMLR, 2022.

[76] Q. Qian, S. Zhu, J. Tang, R. Jin, B. Sun, and H. Li. Robust optimization over multiple domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4739–4746, 2019.

[77] Y. Quan, C. Wang, S. Zhai, M. Fang, and Z. Liu. Enhancing privacy in decentralized min-max optimization: A differentially private approach. In *Proceedings of the Twenty-sixth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pages 371–380, 2025.

[78] M. Ren, W. Zeng, B. Yang, and R. Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pages 4334–4343. PMLR, 2018.

[79] R. Sato, M. Tanaka, and A. Takeda. A gradient method for multilevel optimization. *Advances in Neural Information Processing Systems*, 34:7522–7533, 2021.

[80] A. Shafiei, V. Kungurtsev, and J. Marecek. Trilevel and multilevel optimization using monotone operator theory. *Mathematical Methods of Operations Research*, 99(1):77–114, 2024.

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

[82] Y. Shi, Y. Liu, K. Wei, L. Shen, X. Wang, and D. Tao. Make landscape flatter in differentially private federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24552–24562, 2023.

[83] J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu, and D. Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32, 2019.

[84] B. Song, I. Tsaknakis, C.-Y. Yau, H.-T. Wai, and M. Hong. Distributed optimization for overparameterized problems: Achieving optimal dimension independent communication complexity. *Advances in Neural Information Processing Systems*, 35:6147–6160, 2022.

[85] S. Sra, A. W. Yu, M. Li, and A. Smola. Adadelay: Delay adaptive distributed stochastic optimization. In *Artificial Intelligence and Statistics*, pages 957–965. PMLR, 2016.

[86] S. U. Stich. Local sgd converges fast and communicates little. In *ICLR 2019-International Conference on Learning Representations*, 2019.

[87] E. Suonperä and T. Valkonen. Linearly convergent bilevel optimization with single-step inner methods. *Computational Optimization and Applications*, 87(2):571–610, 2024.

[88] M. Vero, M. Balunovic, D. I. Dimitrov, and M. Vechev. Tableak: Tabular data leakage in federated learning. In *International Conference on Machine Learning*, pages 35051–35083. PMLR, 2023.

[89] D. Wang, X. Pang, J. Hu, S. Yue, and J. Ren. Two-dimensional stackelberg game-based incentive mechanism for differential private federated learning with non-iid data. *IEEE Transactions on Mobile Computing*, 2026.

[90] L. Wang, L. Guo, S. Yang, and X. Shi. Differentially private decentralized optimization with relay communication. *IEEE Transactions on Information Forensics and Security*, 20:724–736, 2024.

[91] N. Wang, Y. Xiao, Y. Chen, N. Zhang, W. Lou, and Y. T. Hou. Squeezing more utility via adaptive clipping on differentially private gradients in federated meta-learning. In *Proceedings of the 38th annual computer security applications conference*, pages 647–657, 2022.

[92] X. Wang, W. Yang, C. Yao, M. Song, and Y. Wan. Revisiting differentially private algorithms for decentralized online learning. In *Forty-second International Conference on Machine Learning*, 2025.

[93] Y. Wang, L. Lin, and J. Chen. Communication-efficient adaptive federated learning. In *International conference on machine learning*, pages 22802–22838. PMLR, 2022.

[94] C. Wei, W. Li, C. Gong, and W. Chen. Dc-sgd: Differentially private sgd with dynamic clipping through gradient norm distribution estimation. *IEEE Transactions on Information Forensics and Security*, 20:4498–4511, 2025.

[95] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE transactions on information forensics and security*, 15:3454–3469, 2020.

[96] C. Wu, F. Wu, T. Qi, Y. Huang, and X. Xie. Fedattack: Effective and covert poisoning attack on federated recommendation via hard sampling. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4164–4172, 2022.

[97] M. Xiang, S. Ioannidis, E. Yeh, C. Joe-Wong, and L. Su. Efficient federated learning against heterogeneous and non-stationary client unavailability. *Advances in Neural Information Processing Systems*, 37:104281–104328, 2024.

[98] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[99] Q. Xiao, S. Lu, and T. Chen. An alternating optimization method for bilevel problems under the polyak-łojasiewicz condition. *Advances in Neural Information Processing Systems*, 36: 63847–63873, 2023.

[100] R. Xie, C. Li, X. Zhou, H. Chen, and Z. Dong. Differentially private federated learning for multitask objective recognition. *IEEE Transactions on Industrial Informatics*, 20(5): 7269–7281, 2024.

[101] C. Yadav and L. Bottou. Cold case: The lost mnist digits. *Advances in neural information processing systems*, 32, 2019.

[102] G. Yan, T. Li, S.-L. Huang, T. Lan, and L. Song. Ac-sgd: Adaptively compressed sgd for communication-efficient distributed learning. *IEEE Journal on Selected Areas in Communications*, 40(9):2678–2693, 2022.

[103] N. Yang and H.-L. Liu. Bilevel evolutionary multi-objective algorithm with multiple lower-level search modes. *IEEE Transactions on Evolutionary Computation*, 2025.

[104] X. Yang, W. Huang, and M. Ye. Dynamic personalized federated learning with adaptive differential privacy. *Advances in Neural Information Processing Systems*, 36:72181–72192, 2023.

[105] Y. Yang, B. Hui, H. Yuan, N. Gong, and Y. Cao. {PrivateFL}: Accurate, differentially private federated learning via personalized data transformation. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 1595–1612, 2023.

[106] Y. Yang, P. Xiao, and K. Ji. Simfbo: Towards simple, flexible and communication-efficient federated bilevel learning. *Advances in Neural Information Processing Systems*, 36:33027–33040, 2023.

[107] Y. Yang, P. Xiao, S. Ma, and K. Ji. First-order federated bilevel learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22029–22037, 2025.

[108] Y. Yang, G. Long, Q. Lu, L. Zhu, and J. Jiang. Bi-level personalization for federated foundation models: A task-vector aggregation approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 27702–27710, 2026.

[109] Y. Yao, T. Edmunds, D. Papageorgiou, and R. Alvarez. Trilevel optimization in power network defense. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(4):712–718, 2007.

[110] J. Yuan, S. Wang, S. Wang, Y. Li, X. Ma, A. Zhou, and M. Xu. Privacy as a resource in differentially private federated learning. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2023.

[111] Y. Yuan, Y. Deng, X. Zhang, Y. Zou, Y. Shi, and D. Yu. Federated bilevel learning against model poisoning attacks. *IEEE Transactions on Networking*, 2025.

[112] X.-Y. Yue, J.-W. Xiao, X.-K. Liu, and Y.-W. Wang. Differentially private linearized admm algorithm for decentralized nonconvex optimization. *IEEE Transactions on Information Forensics and Security*, 2025.

[113] H. Zhang, C. Li, N. Kan, Z. Zheng, W. Dai, J. Zou, and H. Xiong. Improving generalization in federated learning with model-data mutual information regularization: A posterior inference approach. *Advances in Neural Information Processing Systems*, 37:136646–136678, 2024.

[114] L. Zhang, T. Zhu, H. Zhang, P. Xiong, and W. Zhou. Fedrecovery: Differentially private machine unlearning for federated learning frameworks. *IEEE Transactions on Information Forensics and Security*, 18:4732–4746, 2023.

[115] X. Zhang, X. Chen, M. Hong, Z. S. Wu, and J. Yi. Understanding clipping for federated learning: Convergence and client-level differential privacy. In *International Conference on Machine Learning, ICML 2022*, 2022.

[116] X. Zhang, H. Zhao, W. Zhang, S. Cao, H. Sun, and B. Zhang. Fedsw-tsad: Swgan-based federated time series anomaly detection. *Sensors*, 25(13):4014, 2025.

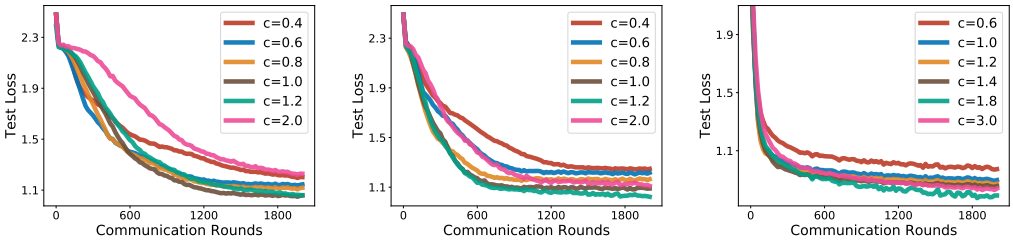
[117] Y. Zhang, R. Dhaipule, C. C. Tan, H. Ling, and H. Gao. Federated stochastic bilevel optimization with fully first-order gradients. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 7047–7055, 2025.

[118] P. Zhao, L. Lai, L. Shen, Q. Li, J. Wu, and Z. Liu. A huber loss minimization approach to mean estimation under user-level differential privacy. *Advances in Neural Information Processing Systems*, 37:130018–130056, 2024.

[119] L. Zhu, Z. Liu, and S. Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

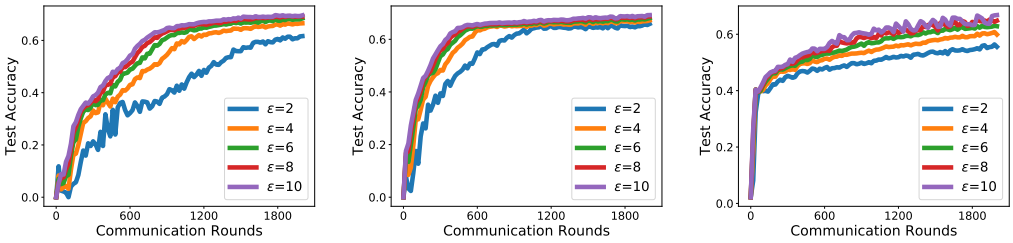
[120] S. Zhu, B. Kong, S. Lu, X. Huang, and K. Yuan. Sparkle: a unified single-loop primal-dual framework for decentralized bilevel optimization. *Advances in Neural Information Processing Systems*, 37:62912–62987, 2024.

[121] G. Zizzo, A. Rawat, M. Sinn, and B. Buesser. Fat: Federated adversarial training. In *Annual Conference on Neural Information Processing Systems*, 2020.



(a) MNIST (b) QMNIST (c) Fashion MNIST

Figure 6: The impact of various clipping thresholds on test loss in (a) MNIST, (b) QMNIST, and (c) Fashion MNIST datasets on DDRO problem with setting $\epsilon = 8$.



(a) MNIST (b) QMNIST (c) Fashion MNIST

Figure 7: The impact of various privacy budget ϵ on test accuracy in (a) MNIST, (b) QMNIST, and (c) Fashion MNIST datasets on DDRO problem.

A Experimental Setting & Additional Results

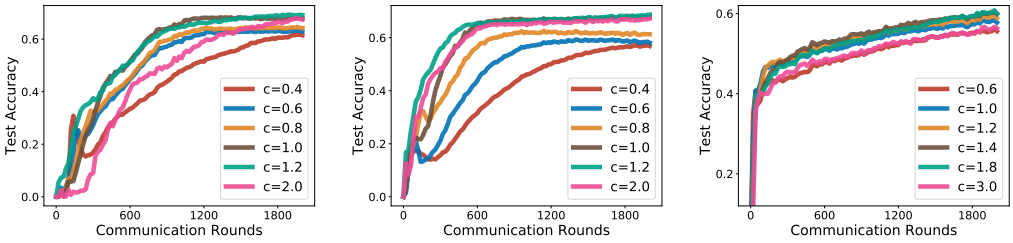
In this section, we provide the experimental settings and additional experimental results in distributed distributionally robust optimization and distributed hyperparameter optimization problems. All experiments were conducted on a server equipped with two NVIDIA 4090 GPUs, and the models were implemented using PyTorch.

A.1 DDRO

While distributed distributionally robust optimization has been widely recognized for its effectiveness in mitigating data heterogeneity across clients [73, 17, 46], distributed systems remain vulnerable to adversarial attacks. To address this dual challenge, we introduce an advanced distributed distributionally robust optimization framework designed not only to handle data heterogeneity but also to defend against adversarial attacks, which can be formulated as a distributed trilevel optimization problem:

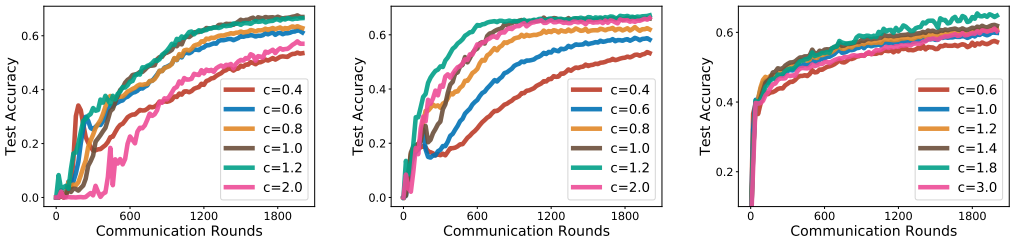
$$\begin{aligned}
&\max \sum_{i=1}^{n} p_i \sum_{(x_j,y_j)\in D_i} \tfrac{1}{|D_i|}\mathcal{L}_p(\boldsymbol{w}, x_j+\gamma_j; y_j)\\
&\text{s.t.}\ \ \boldsymbol{w} = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{n} p_i \sum_{(x_j,y_j)\in D_i} \tfrac{1}{|D_i|}\mathcal{L}_p(\boldsymbol{w}, x_j+\gamma_j; y_j)\\
&\quad \text{s.t.}\ \ \boldsymbol{\gamma} = \arg\max_{\boldsymbol{\gamma}} \sum_{i=1}^{n} \sum_{(x_j,y_j)\in D_i} \tfrac{1}{|D_i|}\mathcal{L}_p(\boldsymbol{w}, x_j+\gamma_j; y_j)\\
&\text{var.} \qquad\quad \mathbf{p}, \boldsymbol{w}, \boldsymbol{\gamma},
\end{aligned} \tag{20}$$

where $\mathbf{p} = [p_1, p_2, \cdots, p_n]$, $\boldsymbol{w}$, and $\boldsymbol{\gamma}$ represent the adversarial distribution, model parameters, and adversarial noise, respectively. For a distributed system of n workers, D_i denotes the local dataset at worker i with a sample size of $|D_i|$, where (x_j, y_j) are the j^{th} data and label. Adopting the configuration from [75], we set $\delta = 1/\sum|D_i|$ in the experiment. In line with previous DDRO studies [46, 76], we evaluate the proposed DPDTO on digit recognition tasks using MNIST [52], QMNIST [101], Fashion-MNIST [98], and USPS [36] datasets. Each worker minimizes the local cross-entropy loss $\mathcal{L}_p$ using a Multilayer Perceptron as the base model. Following the setup of workers and local



(a) MNIST (b) QMNIST (c) Fashion MNIST

Figure 8: The impact of various clipping thresholds with the setting $\epsilon = 4$ on test accuracy in (a) MNIST, (b) QMNIST, and (c) Fashion MNIST datasets on DDRO problem.



(a) MNIST (b) QMNIST (c) Fashion MNIST

Figure 9: The impact of various clipping thresholds with the setting $\epsilon = 8$ on test accuracy in (a) MNIST, (b) QMNIST, and (c) Fashion MNIST datasets on DDRO problem.

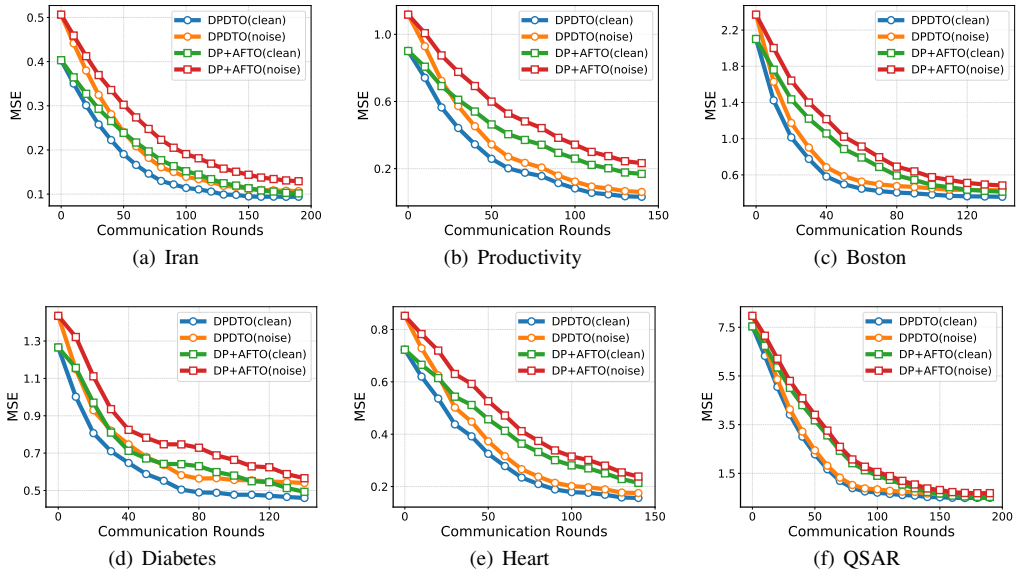
datasets in previous works on distributed distributionally robust optimization [73, 46], we partition the dataset across five workers according to their class labels, with each worker containing two classes of data. To demonstrate two-level robustness, we follow the evaluation standards of [46, 17] by reporting the worst-case worker's defense capabilities against a PGD-7 adversarial attack ($\varepsilon = 0.05$) [69]. Each experiment was independently repeated five times, with both the mean and standard deviation reported in Table 1.

In addition to the results presented in the main text, this section provides several additional experimental results. First, we show the impact of various privacy budget on the test accuracy of the proposed DPDTO in Fig. 7. It can be seen that a relatively larger privacy budget result in better performance. Next, we evaluate the impact of different clipping thresholds on the performance of DPDTO, the experiments are conducted under the setting $\epsilon = 4$ and $\epsilon = 8$. As shown in Figs. 8 and 9, the performance of the proposed method initially improves as the clipping threshold increases within a certain range, since an overly small clipping threshold leads to excessive truncation and introduces additional bias. In addition, setting clipping threshold too large affects the results since more noise is required to be injected to maintain the same level of privacy guarantee [95].

A.2 DRHO

Robust hyperparameter optimization, introduced by [79], provides a principled approach to jointly tuning hyperparameters while simultaneously training a model resilient to perturbations. Following the previous work [42, 44], the distributed robust hyperparameter optimization is considered in the experiment, which can be expressed as

$$\begin{aligned} &\min \sum_{i=1}^{n} \frac{1}{|D_i^{\mathrm{val}}|} \sum_{(\boldsymbol{a}_k, b_k) \in D_i^{\mathrm{val}}} ||b_k - \boldsymbol{a}_k^\top \boldsymbol{\theta}||^2 \\ &\text{s.t. } \boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}'} \sum_{i=1}^{n} \frac{\sum_{(\boldsymbol{a}_k, b_k) \in D_i^{\mathrm{tr}}} ||b_k - (\boldsymbol{a}_k + \boldsymbol{p}'_{i,k})^\top \boldsymbol{\theta}'||^2 + e^{\boldsymbol{\varphi}} ||\boldsymbol{\theta}'||^2}{|D_i^{\mathrm{tr}}|} \\ &\quad \text{s.t. } \mathbf{p} = \arg\max_{\mathbf{p}'} \sum_{i=1}^{n} \frac{\sum_{(\boldsymbol{a}_k, b_k) \in D_i^{\mathrm{tr}}} ||b_k - (\boldsymbol{a}_k + \boldsymbol{p}'_{i,k})^\top \boldsymbol{\theta}'||^2 - r||\boldsymbol{p}'_{i,k}||^2}{|D_i^{\mathrm{tr}}|} \\ &\text{var.} \qquad \boldsymbol{\varphi}, \boldsymbol{\theta}, \mathbf{p}, \end{aligned} \tag{21}$$



(a) Iran (b) Productivity (c) Boston

(d) Diabetes (e) Heart (f) QSAR

Figure 10: Comparison between the proposed DPDTO with the state-of-the-art distributed trilevel optimization method with differential privacy technique, i.e., DP+AFTO, on DRHO problem.

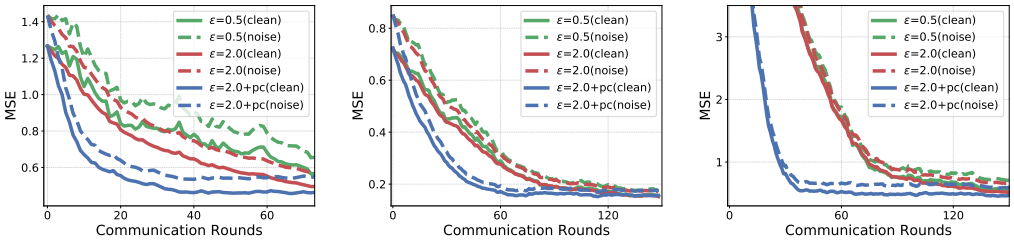
where $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}$ represent the hyperparameters and model parameters, respectively. The term $\mathbf{p}$ indicates the perturbation noise, which is constrained by a penalty coefficient r. Each worker i ($i = 1, \ldots, n$) possesses a training set D_i^{tr} of size $|D_i^{\text{tr}}|$ and a validation set D_i^{val} of size $|D_i^{\text{val}}|$, and $(\boldsymbol{a}_k, b_k)$ stands for the k^{th} input sample and its associated label.

The proposed DPDTO is evaluated using datasets including Diabetes and Boston following [79], and Heart, Iranian Churn, Productivity Prediction, and QSAR Bioconcentration from UCI Machine Learning Repository[2]. Due to space constraints, we could not include all experimental results in the main text. In this section, we present additional results. First, it is seen from Fig. 10 that the proposed DPDTO outperforms the state-of-the-art distributed trilevel optimization method DP+AFTO. This is because: 1) the proposed DPDTO constitutes the first differentially private trilevel optimization framework that provides theoretical guarantees for both convergence and privacy. 2) A theoretically grounded privacy-preserving cutting plane method is developed, which employs a single-level cutting plane approximation instead of the nested-layer strategy adopted in AFTO, thereby effectively mitigating approximation errors. In addition, it is shown in Fig. 11 that 1) the proposed method can be applied in strong privacy requirement, 2) increasing the privacy budget can result in a better performance, and 3) the periodic communication can be used in the proposed method, which can reduce the communication complexity. Moreover, it can be observed from Fig. 12 that relatively increase the clipping threshold can result in better performance. This occurs because an excessively small clipping threshold leads to severe truncation, thereby introducing additional bias.

B Motivating Applications & Threat Model

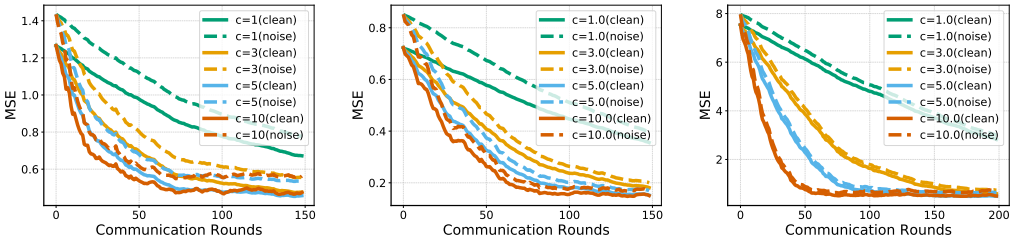
Currently, preserving data privacy has become a critical imperative. For instance, prior work [23] has demonstrated that even when raw datasets are not directly disclosed, publicly accessible trained models remain vulnerable to privacy leakage attacks. Such vulnerabilities pose severe threats to genomic privacy, as adversaries may successfully infer patients' sensitive genetic markers from the shared models. Similar privacy risks also arise in trilevel optimization. Although existing distributed trilevel methods avoid the direct transmission of raw data, the exchanged intermediate information may still expose sensitive private information, rendering these methods vulnerable to privacy leakage attacks. To address this critical vulnerability, we propose DPDTO, the first differentially private trilevel optimization framework. The proposed method is capable of efficiently solving the trilevel

[2]https://archive.ics.uci.edu/



(a) Diabetes (b) Heart (c) QSAR

Figure 11: The impact of various privacy budget ϵ and periodic communication on the performance of DPDTO in (a) Diabetes, (b) Heart, and (c) QSAR datasets on DRHO problem.



(a) Diabetes (b) Heart (c) QSAR

Figure 12: The impact of various clipping thresholds on the performance of DPDTO in (a) Diabetes, (b) Heart, and (c) QSAR datasets on DRHO problem.

optimization problems in a distributed manner while effectively preserving data privacy. In this section, we first present several motivating applications of this work and subsequently discuss the threat model.

B.1 Robust Hyperparameter Optimization in Healthcare

In collaborative healthcare networks, hyperparameter optimization is essential for deploying reliable machine learning models [28, 21]. To ensure model reliability against unpredictable data distribution shifts or adversarial perturbations, robust hyperparameter optimization aims to learn robust machine learning models, which is naturally formulated as a trilevel optimization problem [79]. In such scenarios, sensitive patient data are distributed across hospitals. Although existing distributed trilevel optimization methods can be used without directly transferring raw data, they still face severe risks of privacy leakage. For instance, highly sensitive medical data (e.g., chest X-ray images) can be reconstructed from shared gradients via gradient inversion attacks [29]. This vulnerability motivates us to develop a differentially private trilevel optimization framework in this work.

B.2 Domain Adaptation for Pretraining and Finetuning in Finance

Developing robust predictive models within decentralized financial ecosystems has become a critical trend [4]. To achieve high performance, these systems increasingly depend on the pretrain-finetune paradigm [47]. However, discrepancies between broad pretraining corpora and localized client data may lead to negative transfer. To alleviate negative transfer induced by distribution shifts across regional markets, domain adaptation through data reweighting aims to adaptively adjust the importance of pretraining samples, a process that can be naturally formulated as a trilevel optimization problem [15]. In real-world banking scenarios, proprietary financial records are typically siloed across institutions. While existing distributed trilevel optimization algorithms avoid the direct aggregation of raw data, they may still be vulnerable to potential privacy breaches. Specifically, adversaries can exploit communicated gradient updates to execute inversion attacks, potentially reconstructing highly sensitive tabular attributes, such as income information, credit-related features, and other privacy-critical financial indicators [88]. This security gap motivates us to develop a differentially private trilevel optimization framework.

B.3 Robust OOD Generalization for Time Series Anomaly Detection

In industrial Internet of Things (IoT) and smart grid scenarios, time series anomaly detection plays a pivotal role in enabling early fault diagnosis and ensuring system reliability [116]. To enhance robustness under distribution shifts, trilevel optimization frameworks have been introduced for time series out-of-distribution (OOD) generalization [38]. In practical deployments, however, sensor data are typically stored locally on edge devices within isolated facility networks. Despite avoiding direct raw data sharing, existing distributed trilevel optimization methods may still expose sensitive time series information through intermediate transmissions during training models [71]. Therefore, investigating a privacy-preserving trilevel optimization framework is highly crucial and holds profound practical significance.

B.4 Threat Model

In this work, we aim to preserve privacy when solving trilevel optimization problems. The proposed framework provides robust defenses against a variety of adversarial threat models, including: 1) an honest-but-curious server/master that passively exploits individual workers' local information [51], 2) an untrusted client/worker that passively infers or actively extracts peer workers' local data [105], and 3) an external adversary observing the final model [75]. Following the setting in [75], this work primarily focuses on record-level differential privacy for the joint dataset. However, the proposed framework can be easily adapted to user-level privacy by appropriately adjusting the sensitivities in Eqs. (5), (8), (13), (14), and (15), as discussed in [75].

C Discussion

Relationship with Federated Learning. Standard federated learning is typically formulated as a distributed single-level optimization problem [113, 97], wherein clients and a central server iteratively communicate to optimize a machine learning model. The proposed DPDTO is closely related to federated learning, it can be conceptualized as differentially private federated trilevel learning. This is because the proposed framework also operates within a distributed setting involving client-server/worker-master communication, and the periodic communication mechanisms can be effectively integrated into it (as shown by the results in Figs. 4 and 11). The fundamental distinction lies in that this work addresses a trilevel optimization problem, which is generally more challenging than conventional federated learning. In particular, the constraint in trilevel optimization is itself a bilevel optimization problem [27], and even finding a feasible solution in trilevel optimization is NP-hard. By contrast, traditional federated learning is often formulated as an unconstrained single-level optimization problem or single-level optimization problem with consensus constraints.

Data Heterogeneity. The proposed DPDTO framework is highly effective in tackling the challenge of data heterogeneity. In recent years, numerous studies have demonstrated that the adverse impacts of data heterogeneity in distributed optimization can be mitigated by leveraging bilevel optimization techniques. For instance, [17, 46] effectively alleviate data heterogeneity by formulating a distributionally robust optimization problem via a bilevel optimization framework. Additionally, [19] successfully addresses this issue by applying bilevel optimization driven meta-learning techniques to federated learning. In this paper, we develop a generalized distributed trilevel optimization framework capable of efficiently solving bilevel and trilevel problems with guaranteed data privacy, which can be naturally leveraged to alleviate the effects of data heterogeneity.

Selecting Clipping Thresholds. The clipping thresholds c_x, c_y, and c_z control the trade-off between clipping distortion and the magnitude of privacy noise. In practice, these thresholds can be selected following the DC-SGD-E strategy [94]. Specifically, the gradient-norm distribution at each level is estimated using a differentially private histogram. The corresponding clipping threshold is then chosen to minimize the estimated total error, which accounts for both clipping distortion and the variance of the added privacy noise.

Heterogeneous Privacy Requirements. DPDTO can naturally accommodate heterogeneous privacy requirements by assigning each worker i an individual privacy budget (ϵ_i, δ_i) and worker-specific noise multipliers $\sigma_{x,i}, \sigma_{y,i}, \sigma_{z,i}$. Specifically, with $q_{1,i} = m_i/M_i$ and $q_{2,i} = \widetilde{m}_i/M_i$, the noise

Table 3: Notations and their meanings in this work.

Notations	Meaning
n, $[n]$	Number of workers and worker index set $\{1, \ldots, n\}$
$f_{b,i}(\cdot)$	Level-b local objective function at worker i, $b \in [3]$, $i \in [n]$
$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$	First-, second-, and third-level optimization variables
d_1, d_2, d_3	Dimensions of $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$, respectively
t, T, K	Current iteration, total iterations, and period of value-function/cutting plane updates
M, m, $\tilde{m}$	Total number of data records and mini-batch sizes
$\xi_{i,k,t}$	Random variable
ϑ, λ	Lagrange multipliers
$\mathcal{C}^t$	Feasible-region relaxation formed by the cutting planes at iteration t
$\mathbb{C}_{\mathrm{DP}}^{\mathrm{new}}$	Newly generated differentially private cutting plane
$\boldsymbol{a}_{x,l}, \boldsymbol{a}_{y,l}, \boldsymbol{a}_{z,l}$	Quadratic coefficients of the l^{th} cutting plane
$\boldsymbol{b}_{x,l}, \boldsymbol{b}_{y,l}, \boldsymbol{b}_{z,l}, c_l$	Linear coefficients and scalar offset of the l^{th} cutting plane
$o_l(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$	Constraint-residual function of the l^{th} cutting plane
$L_\lambda(\cdot)$, $L_{\lambda,i}(\cdot)$	Global and local Lagrangian functions of the cutting plane relaxation
$\mathrm{clip}_{c_{\boldsymbol{u}}}(\cdot)$, $c_{\boldsymbol{u}}$	Gradient-clipping operator and clipping threshold for $\boldsymbol{u} \in \{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}\}$
$\boldsymbol{g}_{\boldsymbol{u}}^{i,t}, \Delta_{\boldsymbol{u}}^{i,t}, \overline{\boldsymbol{g}}_{\boldsymbol{u}}^{t}$	Clipped local gradient, privatized local update, and worker average for $\boldsymbol{u} \in \{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}\}$
$\tilde{\Delta}_{\boldsymbol{y}}^{i,t}, \tilde{\Delta}_{\boldsymbol{z}}^{i,t}$	Privatized local gradients used in cutting plane and value-function updates
$\eta_{\boldsymbol{x}}, \eta_{\boldsymbol{y}}, \eta_{\boldsymbol{z}}$	Step sizes for updating the three level variables
$\boldsymbol{v}_{\boldsymbol{u}}^{i,t}, \tilde{\boldsymbol{v}}_{\boldsymbol{u}}^{i,t}, \sigma_{\boldsymbol{u}}$	Gaussian noise vectors and corresponding noise multiplier for $\boldsymbol{u} \in \{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}\}$
HVX_i^t, HVZ_i^t	Noisy Hessian-vector products for cutting plane construction
ϵ, δ	Privacy budget and failure probability of (ϵ, δ)-differential privacy
$q_1 = \frac{m}{M}, q_2 = \frac{\tilde{m}}{M}$	Sampling rates for variable updates and value-function/cutting plane updates
L	Parameter in L-smoothness
$D_{\boldsymbol{x}}, D_{\boldsymbol{y}}, D_{\boldsymbol{z}}$	Variable bounds in the iterative process
$G_{\boldsymbol{x}}, G_{\boldsymbol{y}}, G_{\boldsymbol{z}}$	Constants related to gradient boundedness G_f

multipliers can be calibrated as

$$\min\{\sigma_{x,i}, \sigma_{y,i}, \sigma_{z,i}\} \geq \sqrt{3q_{1,i}^2 T + 4q_{2,i}^2 \lfloor T/K \rfloor} \frac{2\sqrt{\beta_0 \log(1/\delta_i)}}{\epsilon_i}.$$

Consequently, the complete transcript associated with worker i satisfies (ϵ_i, δ_i)-DP. Since workers hold disjoint local datasets, parallel composition provides heterogeneous worker-level privacy guarantees, a uniform system-level guarantee can also be stated as $(\max_i \epsilon_i, \max_i \delta_i)$-DP.

D Proof of Theorem 4.3

Theorem 4.3 (Asymptotic Convergence of the Cumulative Relaxation) Let $\boldsymbol{\mathcal{C}}^t$ denote the cumulative feasible region by retaining all generated privacy-preserving cutting planes at t^{th} iteration. It follows that $\boldsymbol{\mathcal{C}}^t$ is a relaxation of the original feasible region when $\phi(\cdot)$ is L-smooth, and it converges as more cutting planes are added:

$$\begin{aligned} &\boldsymbol{\mathcal{H}} = \{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} | \phi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \leq 0\} \subseteq \boldsymbol{\mathcal{C}}^t = \{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} | \mathbb{E}[o_l(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})] \leq 0, \forall l\}, \\ &\boldsymbol{\mathcal{H}} \subseteq \lim_{t \to \infty} \downarrow \boldsymbol{\mathcal{C}}^t \subseteq \cdots \subseteq \boldsymbol{\mathcal{C}}^t \subseteq \boldsymbol{\mathcal{C}}^{t-1} \subseteq \cdots \subseteq \boldsymbol{\mathcal{C}}^0. \end{aligned} \tag{22}$$

Proof:

Defining functions $V(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)$ and $\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)$ as:

$$V(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t) = \boldsymbol{y}^t - \boldsymbol{y}_0 + \frac{\eta_{\boldsymbol{y}}}{n\tilde{m}} \sum_{i=1}^{n} \sum_{k=1}^{\tilde{m}} \nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k}), \tag{23}$$

$$
\begin{aligned}
&\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t) \\
&= \hat{V}(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t) - V(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t) \\
&= \frac{\eta_{\boldsymbol{y}}}{n\tilde{m}} \sum_{i=1}^{n} \sum_{k=1}^{\tilde{m}} \Big(\mathrm{clip}_{c_{\boldsymbol{y}}}(\nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})) - \nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k}) \Big) + \frac{2\eta_{\boldsymbol{y}} c_{\boldsymbol{y}}}{n\tilde{m}} \sum_{i=1}^{n} \tilde{v}_{\boldsymbol{y}}^{i,t}.
\end{aligned}
\tag{24}
$$

Given that the lower-level optimization problem serves as a soft constraint on the upper-level objective, it is common practice to approximate the lower-level solution using a single gradient descent step [83, 20, 61, 78, 87]. Consequently, we have

$$
\begin{aligned}
&\mathbb{E}[\hat{\phi}(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t) - \phi(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)] \\
&= \mathbb{E}\left[||V(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t) + \delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2 - ||V(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2 \right] \\
&= \mathbb{E}\left[2\langle V(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t), \delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t) \rangle + ||\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2 \right] \\
&\leq 2\mathbb{E}\left[||V(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)|| \cdot ||\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)|| \right] + \mathbb{E}[||\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2].
\end{aligned}
\tag{25}
$$

In addition, based on the triangle inequality and Assumption 4.2, we have:

$$
\begin{aligned}
&\mathbb{E}[||\nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})||] \\
&= \mathbb{E}[||\nabla_{\boldsymbol{y}} f_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k}) + \vartheta \nabla_{\boldsymbol{y}} f_{3,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k}) - \vartheta \nabla_{\boldsymbol{y}} f_{3,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^*, \xi_{i,k})||] \\
&\leq \mathbb{E}[||\nabla_{\boldsymbol{y}} f_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})|| + \vartheta ||\nabla_{\boldsymbol{y}} f_{3,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})|| + \vartheta ||\nabla_{\boldsymbol{y}} f_{3,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^*, \xi_{i,k})||] \\
&\leq (1 + 2\vartheta)\sqrt{G_f}.
\end{aligned}
\tag{26}
$$

Applying the triangle inequality yields

$$
\begin{aligned}
&\mathbb{E}[||V(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||] \\
&\leq \mathbb{E}[||\boldsymbol{y}^t - \boldsymbol{y}_0||] + \frac{\eta_{\boldsymbol{y}}}{n\tilde{m}} \sum_{i=1}^{n} \sum_{k=1}^{\tilde{m}} \mathbb{E}[||\nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})||] \\
&\leq 2\sqrt{D_{\boldsymbol{y}}} + \eta_{\boldsymbol{y}}(1 + 2\vartheta)\sqrt{G_f}.
\end{aligned}
\tag{27}
$$

Furthermore, it follows that

$$
\begin{aligned}
&\mathbb{E}[||\mathrm{clip}_{c_{\boldsymbol{y}}}(\nabla_{\boldsymbol{y}} g_{2,i}(x, \boldsymbol{y}_0, z, \xi_{i,k})) - \nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})||^2] \\
&\leq \mathbb{E}[||\nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k}) - \frac{c_{\boldsymbol{y}}}{||\nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})||} \nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})||^2] \\
&= \mathbb{E}[||\nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})||^2 ||1 - \frac{c_{\boldsymbol{y}}}{||\nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})||}||^2] \\
&\leq \mathbb{E}[||\nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})||^2].
\end{aligned}
\tag{28}
$$

By combining $\mathbb{E}[\tilde{v}_{\boldsymbol{y}}^{i,t}] = 0$ with Eqs. (24)-(28), we can get:

$$
\begin{aligned}
&\mathbb{E}[||\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2] \\
&= \frac{\eta_{\boldsymbol{y}}^2}{n^2\tilde{m}^2}\mathbb{E}[||\sum_{i=1}^{n}\sum_{k=1}^{\tilde{m}} \mathrm{clip}_{c_{\boldsymbol{y}}}(\nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})) - \nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})||^2] \\
&+ \frac{4\eta_{\boldsymbol{y}}^2 {c_{\boldsymbol{y}}}^2}{n^2\tilde{m}^2}\mathbb{E}[||\sum_{i=1}^{n}\tilde{v}_{\boldsymbol{y}}^{i,t}||^2] \\
&\leq \frac{\eta_{\boldsymbol{y}}^2}{nm}\sum_{i=1}^{n}\sum_{k=1}^{\tilde{m}} \mathbb{E}[||\mathrm{clip}_{c_{\boldsymbol{y}}}(\nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})) - \nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})||^2] \\
&+ \frac{4\eta_{\boldsymbol{y}}^2 {c_{\boldsymbol{y}}}^2}{n^2\tilde{m}^2}\mathbb{E}[||\sum_{i=1}^{n}\tilde{v}_{\boldsymbol{y}}^{i,t}||^2] \\
&\leq \frac{\eta_{\boldsymbol{y}}^2}{n\tilde{m}}\sum_{i=1}^{n}\sum_{k=1}^{\tilde{m}} \mathbb{E}[||\nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})||^2] + \frac{4\eta_{\boldsymbol{y}}^2 {c_{\boldsymbol{y}}}^2}{n^2\tilde{m}^2}\sum_{i=1}^{n}\mathbb{E}[||\tilde{v}_{\boldsymbol{y}}^{i,t}||^2] \\
&\leq \eta_{\boldsymbol{y}}^2\left((1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}\right).
\end{aligned}
\tag{29}
$$

Furthermore, it follows from Eq. (29) that $\mathbb{E}[||\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||] \leq \eta_{\boldsymbol{y}}\sqrt{(1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}}$. By substituting this into Eqs. (25) and (27), we arrive at

$$
\begin{aligned}
&\mathbb{E}[\hat{\phi}(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t) - \phi(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)] \\
&= \mathbb{E}\left[||V(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t) + \delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2 - ||V(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2\right] \\
&= \mathbb{E}\left[2\langle V(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t), \delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)\rangle + ||\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2\right] \\
&\leq \mathbb{E}\left[2||V(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)|| \cdot ||\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)|| + ||\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2\right] \\
&\leq 2\eta_{\boldsymbol{y}}\sqrt{(1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}}\left(2\sqrt{D_{\boldsymbol{y}}} + \eta_{\boldsymbol{y}} L(1+2\vartheta)\sqrt{G_f}\right) + \eta_{\boldsymbol{y}}^2\left((1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}\right) \\
&= 4\eta_{\boldsymbol{y}}\sqrt{(1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}}\sqrt{D_{\boldsymbol{y}}} \\
&+ \eta_{\boldsymbol{y}}^2\left(2L(1+2\vartheta)\sqrt{G_f}\sqrt{(1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}} + (1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}\right).
\end{aligned}
\tag{30}
$$

Then, we define $J_V(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}), J_{\hat{V}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}), J_\delta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ as:

$$
J_V(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \left[\frac{\partial V(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})}{\partial x}, \frac{\partial V(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})}{\partial y}, \frac{\partial V(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})}{\partial z}\right], \tag{31}
$$

$$
J_{\hat{V}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \left[\frac{\partial \hat{V}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})}{\partial x}, \frac{\partial \hat{V}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})}{\partial y}, \frac{\partial \hat{V}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})}{\partial z}\right], \tag{32}
$$

$$
J_\delta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = J_V(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) - J_{\hat{V}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \left[\frac{\partial \delta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})}{\partial x}, \frac{\partial \delta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})}{\partial y}, \frac{\partial \delta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})}{\partial z}\right]. \tag{33}
$$

Applying the triangle inequality in conjunction with the L-smoothness property, we obtain:

$$\begin{aligned}
&\mathbb{E}[||J_V(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||] \\
&\leq \mathbb{E}\left[||\frac{\partial V(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)}{\partial x}||\right] + \mathbb{E}\left[||\frac{\partial V(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)}{\partial y}||\right] + \mathbb{E}\left[||\frac{\partial V(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)}{\partial z}||\right] \\
&\leq \frac{\eta_{\boldsymbol{y}}}{n\tilde{m}} \sum_{i=1}^{n} \sum_{k=1}^{\tilde{m}} \mathbb{E}[||\nabla^2_{xy} g_{2,i}(\boldsymbol{x}^{t+1}, \boldsymbol{y}_0, \boldsymbol{z}^{t+1}, \xi_k)||] + 1 \\
&+ \frac{\eta_{\boldsymbol{y}}}{n\tilde{m}} \sum_{i=1}^{n} \sum_{k=1}^{\tilde{m}} \mathbb{E}[||\nabla^2_{zy} g_{2,i}(\boldsymbol{x}^{t+1}, \boldsymbol{y}_0, \boldsymbol{z}^{t+1}, \xi_k)||] \\
&= 1 + 2\eta_{\boldsymbol{y}} L,
\end{aligned} \tag{34}$$

and we have that

$$\begin{aligned}
&\frac{\partial \delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)}{\partial \boldsymbol{x}} \\
&= \frac{\eta_{\boldsymbol{y}}}{n\tilde{m}} \sum_{i=1}^{n} \sum_{k=1}^{\tilde{m}} \frac{\partial \left(\text{clip}_{c_{\boldsymbol{y}}}(\nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})) - \nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k}) \right)}{\partial \boldsymbol{x}} \\
&= \frac{\eta_{\boldsymbol{y}}}{n\tilde{m}} \sum_{i=1}^{n} \sum_{k=1}^{\tilde{m}} \left(J_{\text{clip}}(\nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})) \cdot \nabla^2_{xy} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k}) \right) \\
&- \frac{\eta_{\boldsymbol{y}}}{n\tilde{m}} \sum_{i=1}^{n} I \cdot \nabla^2_{xy} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k}) \\
&= \frac{\eta_{\boldsymbol{y}}}{n\tilde{m}} \sum_{i=1}^{n} \sum_{k=1}^{\tilde{m}} \left((J_{\text{clip}}(\nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})) - I) \cdot \nabla^2_{xy} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k}) \right).
\end{aligned} \tag{35}$$

Based on triangle inequality, we can obtain the following inequality:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{\partial \delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)}{\partial \boldsymbol{x}}\right\|\right] \\
&\leq \frac{\eta_{\boldsymbol{y}}}{n\tilde{m}} \sum_{i=1}^{n} \sum_{k=1}^{\tilde{m}} \mathbb{E}[||J_{\text{clip}}(\nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^{t+1}, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})) - I|| \cdot ||\nabla^2_{xy} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})||].
\end{aligned} \tag{36}$$

Combining the triangle inequality with the boundedness of the clipping operator ($J_{\text{clip}}(\cdot) \leq 1$), it follows that:

$$\begin{aligned}
&\mathbb{E}[||J_{\text{clip}}(\nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})) - I||] \\
&\leq \mathbb{E}[||J_{\text{clip}}(\nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k}))|| + ||I||] \\
&\leq 2.
\end{aligned} \tag{37}$$

Integrating the L-smoothness property with Eqs. (36) and (37), we obtain:

$$\mathbb{E}\left[\left\|\frac{\partial \delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)}{\partial \boldsymbol{x}}\right\|\right] \leq 2\eta_{\boldsymbol{y}} L. \tag{38}$$

Likewise, it can be derived that

$$\begin{aligned}
&\mathbb{E}[||\frac{\partial \delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)}{\partial \boldsymbol{z}}||] \\
&= \mathbb{E}[||\frac{\eta_{\boldsymbol{y}}}{n\tilde{m}} \sum_{i=1}^{n} \sum_{k=1}^{\tilde{m}} (J_{\text{clip}}(\nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})) - I) \cdot \nabla^2_{zy} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})||] \\
&\leq \frac{\eta_{\boldsymbol{y}}}{n\tilde{m}} \sum_{i=1}^{n} \sum_{k=1}^{\tilde{m}} \mathbb{E}[||J_{\text{clip}}(\nabla_{\boldsymbol{y}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})) - I|| \cdot ||\nabla^2_{zy} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k})||] \\
&\leq 2\eta_{\boldsymbol{y}} L.
\end{aligned} \tag{39}$$

According to Eq. (36), Eq. (39), and $\frac{\partial \delta(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)}{\partial \boldsymbol{y}} = \mathbf{0}$, we have:

$$\begin{aligned}
&\mathbb{E}[||J_\delta(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)||] \\
&\leq \mathbb{E}\left[\left\|\frac{\partial \delta(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)}{\partial \boldsymbol{x}}\right\| + \left\|\frac{\partial \delta(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)}{\partial \boldsymbol{y}}\right\| + \left\|\frac{\partial \delta(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)}{\partial \boldsymbol{z}}\right\|\right] \\
&\leq 4\eta_{\boldsymbol{y}} L.
\end{aligned} \tag{40}$$

Based on the definition of $\nabla\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)$ and $\nabla\phi(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)$, we have:

$$\begin{aligned}
&\nabla\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t) - \nabla\phi(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t) \\
&= 2J_V(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)^\top \delta(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t) + 2J_\delta(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)^\top V(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t) \\
&+2J_\delta(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)^\top \delta(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t).
\end{aligned} \tag{41}$$

By combining Eq. (41) with the triangle inequality, we can get:

$$\begin{aligned}
&||\nabla\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t) - \nabla\phi(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)|| \\
&\leq 2||J_V(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)^\top \delta(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)|| + 2||J_\delta(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)^\top V(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)|| \\
&+2||J_\delta(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)^\top \delta(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)|| \\
&\leq 2||J_V(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)|| \cdot ||\delta(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)|| + 2||J_\delta(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)|| \cdot ||V(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)|| \\
&+2||J_\delta(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)|| \cdot ||\delta(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)||.
\end{aligned} \tag{42}$$

Hence, we arrive at the following bound:

$$\begin{aligned}
&||\nabla\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t) - \nabla\phi(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)|| \\
&\leq \left(2\eta_{\boldsymbol{y}} + 4\eta_{\boldsymbol{y}}^2 L\right) \cdot \sqrt{(1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}} + 8\eta_{\boldsymbol{y}} L\left(2\sqrt{D_{\boldsymbol{y}}} + \eta_{\boldsymbol{y}}(1+2\vartheta)\sqrt{G_f}\right) \\
&+ 8\eta_{\boldsymbol{y}}^2 L\sqrt{(1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}} \\
&= \eta_{\boldsymbol{y}}\left(2\sqrt{(1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}} + 16L\sqrt{D_{\boldsymbol{y}}}\right) \\
&+ \eta_{\boldsymbol{y}}^2\left(8L(1+2\vartheta)\sqrt{G_f} + 12L\sqrt{(1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}}\right).
\end{aligned} \tag{43}$$

Consequently, it follows that:

$$\begin{aligned}
&\begin{bmatrix} \mathbb{E}[\nabla_{\boldsymbol{x}}\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t) - \nabla_{\boldsymbol{x}}\phi(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)] \\ \mathbb{E}[\nabla_{\boldsymbol{y}}\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t) - \nabla_{\boldsymbol{y}}\phi(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)] \\ \mathbb{E}[\nabla_{\boldsymbol{z}}\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t) - \nabla_{\boldsymbol{z}}\phi(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)] \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{x}-\boldsymbol{x}^t \\ \boldsymbol{y}-\boldsymbol{y}^t \\ \boldsymbol{z}-\boldsymbol{z}^t \end{bmatrix} \\
&\leq ||\nabla\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t) - \nabla\phi(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)|| \cdot (||\boldsymbol{x}-\boldsymbol{x}^t|| + ||\boldsymbol{y}-\boldsymbol{y}^t|| + ||\boldsymbol{z}-\boldsymbol{z}^t||) \\
&\leq \eta_{\boldsymbol{y}}\left(4\sqrt{(1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}} + 32L\sqrt{D_{\boldsymbol{y}}}\right)\left(\sqrt{D_{\boldsymbol{x}}} + \sqrt{D_{\boldsymbol{y}}} + \sqrt{D_{\boldsymbol{z}}}\right) \\
&+ \eta_{\boldsymbol{y}}^2\left(16L(1+2\vartheta)\sqrt{G_f} + 24L\sqrt{(1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}}\right)\left(\sqrt{D_{\boldsymbol{x}}} + \sqrt{D_{\boldsymbol{y}}} + \sqrt{D_{\boldsymbol{z}}}\right).
\end{aligned} \tag{44}$$

For a function satisfying the L-smoothness property, it follows that:

$$\begin{aligned}
&\phi(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t) + \begin{bmatrix} \nabla_{\boldsymbol{x}}\phi(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t) \\ \nabla_{\boldsymbol{y}}\phi(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t) \\ \nabla_{\boldsymbol{z}}\phi(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t) \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{x}-\boldsymbol{x}^t \\ \boldsymbol{y}-\boldsymbol{y}^t \\ \boldsymbol{z}-\boldsymbol{z}^t \end{bmatrix} \\
&- \frac{L}{2}||\boldsymbol{x}-\boldsymbol{x}^t||^2 - \frac{L}{2}||\boldsymbol{y}-\boldsymbol{y}^t||^2 - \frac{L}{2}||\boldsymbol{z}-\boldsymbol{z}^t||^2 \leq \phi(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}).
\end{aligned} \tag{45}$$

Based on Eq. (45), we can get

$$
\begin{aligned}
&\mathbb{E}[\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)] + \begin{bmatrix} \mathbb{E}[\nabla_{\boldsymbol{x}}\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)] \\ \mathbb{E}[\nabla_{\boldsymbol{y}}\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)] \\ \mathbb{E}[\nabla_{\boldsymbol{z}}\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)] \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{x}-\boldsymbol{x}^t \\ \boldsymbol{y}-\boldsymbol{y}^t \\ \boldsymbol{z}-\boldsymbol{z}^t \end{bmatrix} \\
&- \frac{L}{2}||\boldsymbol{x}-\boldsymbol{x}^t||^2 - \frac{L}{2}||\boldsymbol{y}-\boldsymbol{y}^t||^2 - \frac{L}{2}||\boldsymbol{z}-\boldsymbol{z}^t||^2 \\
&\leq \phi(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) + \mathbb{E}[\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t) - \phi(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)] \\
&+ \begin{bmatrix} \mathbb{E}[\nabla_{\boldsymbol{x}}\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t) - \nabla_{\boldsymbol{x}}\phi(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)] \\ \mathbb{E}[\nabla_{\boldsymbol{y}}\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t) - \nabla_{\boldsymbol{y}}\phi(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)] \\ \mathbb{E}[\nabla_{\boldsymbol{z}}\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t) - \nabla_{\boldsymbol{z}}\phi(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)] \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{x}-\boldsymbol{x}^t \\ \boldsymbol{y}-\boldsymbol{y}^t \\ \boldsymbol{z}-\boldsymbol{z}^t \end{bmatrix} \\
&\leq \phi(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) + 2\eta_{\boldsymbol{y}}\sqrt{(1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}}\,2\sqrt{D_{\boldsymbol{y}}} \\
&+ \eta_{\boldsymbol{y}}^2\left(2G\sqrt{(1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}} + (1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}\right) \\
&+ \eta_{\boldsymbol{y}}\left(4\sqrt{(1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}} + 32L\sqrt{D_{\boldsymbol{y}}}\right)\left(\sqrt{D_{\boldsymbol{x}}} + \sqrt{D_{\boldsymbol{y}}} + \sqrt{D_{\boldsymbol{z}}}\right) \\
&+ \eta_{\boldsymbol{y}}^2\left(16L(1+2\vartheta)\sqrt{G_f} + 24L\sqrt{(1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}}\right)\left(\sqrt{D_{\boldsymbol{x}}} + \sqrt{D_{\boldsymbol{y}}} + \sqrt{D_{\boldsymbol{z}}}\right).
\end{aligned}
\tag{46}
$$

Thus, for any point belongs to the original feasible region $\boldsymbol{\mathcal{H}} = \{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}|\phi(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) \leq 0\}$, it always satisfied that

$$
\begin{aligned}
&\mathbb{E}[\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)] + \begin{bmatrix} \mathbb{E}[\nabla_{\boldsymbol{x}}\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)] \\ \mathbb{E}[\nabla_{\boldsymbol{y}}\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)] \\ \mathbb{E}[\nabla_{\boldsymbol{z}}\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)] \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{x}-\boldsymbol{x}^t \\ \boldsymbol{y}-\boldsymbol{y}^t \\ \boldsymbol{z}-\boldsymbol{z}^t \end{bmatrix} \\
&- \frac{L}{2}||\boldsymbol{x}-\boldsymbol{x}^t||^2 - \frac{L}{2}||\boldsymbol{y}-\boldsymbol{y}^t||^2 - \frac{L}{2}||\boldsymbol{z}-\boldsymbol{z}^t||^2 \\
&\leq 4\eta_{\boldsymbol{y}}\sqrt{(1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}}\sqrt{D_{\boldsymbol{y}}} \\
&+ \eta_{\boldsymbol{y}}^2\left(2(1+2\vartheta)\sqrt{G_f}\sqrt{(1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}} + (1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}\right) \\
&+ \eta_{\boldsymbol{y}}\left(2\sqrt{(1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}} + 16L\sqrt{D_{\boldsymbol{y}}}\right)\left(\sqrt{D_{\boldsymbol{x}}} + \sqrt{D_{\boldsymbol{y}}} + \sqrt{D_{\boldsymbol{z}}}\right) \\
&+ \eta_{\boldsymbol{y}}^2\left(16L(1+2\vartheta)\sqrt{G_f} + 24L\sqrt{(1+2\vartheta)^2 G_f + \frac{4{c_{\boldsymbol{y}}}^2\sigma_{\boldsymbol{y}}^2 d_2}{n\tilde{m}^2}}\right)\left(\sqrt{D_{\boldsymbol{x}}} + \sqrt{D_{\boldsymbol{y}}} + \sqrt{D_{\boldsymbol{z}}}\right).
\end{aligned}
\tag{47}
$$

Moreover, as $\mathbb{E}[\frac{1}{n}\sum_{i=1}^n HVX_i^t] = \nabla_{\boldsymbol{x}}\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)$, $2\hat{V}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t) = \nabla_{\boldsymbol{y}}\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)$, $\mathbb{E}[\frac{1}{n}\sum_{i=1}^n HVZ_i^t] = \nabla_{\boldsymbol{z}}\hat{\phi}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)$, combining with Eqs. (10) and (47), we can conclude that the proposed privacy-preserving cutting planes can form a relaxation for the original feasible region: $\boldsymbol{\mathcal{H}} = \{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}|\phi(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) \leq 0\} \subseteq \boldsymbol{\mathcal{C}}^t = \{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}|\mathbb{E}[o_l(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})] \leq 0, \forall l\}$. Moreover, we can derive that with additional cutting planes are added, the relaxation is convergent $\boldsymbol{\mathcal{H}} \subseteq \lim_{t\to\infty}\downarrow\boldsymbol{\mathcal{C}}^t \subseteq \cdots \subseteq \boldsymbol{\mathcal{C}}^t \subseteq \boldsymbol{\mathcal{C}}^{t-1} \subseteq \cdots \subseteq \boldsymbol{\mathcal{C}}^0$.

□

Remark D.1 *In conventional cutting plane methods (e.g., [41, 65, 63, 38, 42]), the transmission of local information, such as gradients, during the generation of cutting planes poses a potential risk of data privacy leakage. The cutting plane method proposed in this paper provides data privacy guarantees. First, during the computation of $\hat{\phi}(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)$, we transmit the local information $\tilde{\Delta}_{\boldsymbol{y}}^{i,t}$ protected by differential privacy, thereby ensuring data privacy at this stage. Furthermore, we consider the potential information leakage associated with the transmission of Hessian-vector products [66, 24]. By leveraging the L-smoothness, we establish that the second-order information is bounded: $||\frac{1}{\tilde{m}}\sum_{k=1}^{\tilde{m}} \nabla^2_{\boldsymbol{xy}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k,t}) \cdot \boldsymbol{w}_{i,k}^t|| \leq \frac{1}{\tilde{m}}\sum_{k=1}^{\tilde{m}} ||\nabla^2_{\boldsymbol{xy}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k,t})|| \cdot ||\boldsymbol{w}_{i,k}^t|| \leq L||\bar{\Delta}^t||$ and $||\frac{1}{\tilde{m}}\sum_{k=1}^{\tilde{m}} \nabla^2_{\boldsymbol{zy}} g_{2,i}(\boldsymbol{x}^t, \boldsymbol{y}_0, \boldsymbol{z}^t, \xi_{i,k,t}) \cdot \boldsymbol{w}_{i,k}^t|| \leq L||\bar{\Delta}^t||$, the sensitivity can be expressed as $\frac{2L||\bar{\Delta}^t||}{\tilde{m}}$. Consequently, Gaussian noise is injected during this process to further safeguard data privacy. Meanwhile, we analytically demonstrate in Theorem 4.3 that the proposed method constitutes a convergent relaxation. To the best of our knowledge, our proposed method is the first cutting plane approach equipped with privacy guarantees. It is worth noting that a naive combination of differential privacy techniques with existing cutting plane methods fails to provide any theoretical guarantees (e.g., valid feasible region relaxation). This is primarily because the underlying process is highly complex, involving operations such as clipping, noise injection, and inter-node communication.*

E Proof of Theorem 4.4

Theorem 4.4 (Non-Asymptotic Convergence of DPDTO) Under Assumptions 4.1 and 4.2, let step-sizes $\eta_{\boldsymbol{x}} = \eta_{\boldsymbol{y}} = \eta_{\boldsymbol{z}} = \frac{1}{\sqrt{T}}$, we can derive:

$$\begin{aligned}
&\textstyle\sum_{t=0}^{T-1} \frac{||\nabla L_\lambda(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2}{T} \leq \frac{(4L_\lambda(\boldsymbol{x}_0, \boldsymbol{y}_0, \boldsymbol{z}_0) - f^*)}{\sqrt{T}} + \left(\frac{2G_{\boldsymbol{x}}^4}{c_{\boldsymbol{x}}^2} + \frac{4G_{\boldsymbol{y}}^4}{c_{\boldsymbol{y}}^2} + \frac{4G_{\boldsymbol{z}}^4}{c_{\boldsymbol{z}}^2}\right) \\
&+ \frac{2\hat{L}^2(c_{\boldsymbol{x}}^2 + c_{\boldsymbol{y}}^2 + c_{\boldsymbol{z}}^2)}{\sqrt{T}} + \frac{8\hat{L}^2}{\sqrt{T}}\left(\frac{d_1 c_{\boldsymbol{x}}^2 \sigma_{\boldsymbol{x}}^2}{nm^2} + \frac{d_2 c_{\boldsymbol{y}}^2 \sigma_{\boldsymbol{y}}^2}{nm^2} + \frac{d_3 c_{\boldsymbol{z}}^2 \sigma_{\boldsymbol{z}}^2}{nm^2}\right) \\
&+ \frac{\hat{L}^2}{T}\left(18c_{\boldsymbol{x}}^2 + 12c_{\boldsymbol{y}}^2 + \frac{24d_1 c_{\boldsymbol{x}}^2 \sigma_{\boldsymbol{x}}^2}{nm^2} + \frac{16d_2 c_{\boldsymbol{y}}^2 \sigma_{\boldsymbol{y}}^2}{nm^2} + \frac{48d_1 c_{\boldsymbol{x}}^2 \sigma_{\boldsymbol{x}}^2}{m^2} + \frac{32d_2 c_{\boldsymbol{y}}^2 \sigma_{\boldsymbol{y}}^2}{m^2}\right),
\end{aligned} \tag{48}$$

where $c_{\boldsymbol{x}}, c_{\boldsymbol{y}}, c_{\boldsymbol{z}}$ are clipping thresholds, d_1, d_2, d_3 are the dimensions of each level variables, $f^* = \min L_\lambda(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, $\hat{L}$ is a constant related to L-smoothness, $G_{\boldsymbol{x}}, G_{\boldsymbol{y}}, G_{\boldsymbol{z}}$ are the constants related to the gradient boundedness G_f, m is the batch size, n is the number of workers in distributed systems.

Proof:

To facilitate the proof of Theorem 4.4, we first provide three key lemmas as follows.

Lemma E.1 *In $(t+1)^{\text{th}}$ iteration, the discrepancy between the local and global variables is upper bounded as follows:*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[||\boldsymbol{x}^{i,t+1} - \boldsymbol{x}^{t+1}||^2] \leq \eta_{\boldsymbol{x}}^2 \left(c_{\boldsymbol{x}}^2 + \frac{d_1 c_{\boldsymbol{x}}^2 \sigma_{\boldsymbol{x}}^2}{m^2}\right), \tag{49}$$

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[||\boldsymbol{y}^{i,t+1} - \boldsymbol{y}^{t+1}||^2] \leq \eta_{\boldsymbol{y}}^2 \left(c_{\boldsymbol{y}}^2 + \frac{d_2 c_{\boldsymbol{y}}^2 \sigma_{\boldsymbol{y}}^2}{m^2}\right). \tag{50}$$

Proof of Lemma E.1:

First, according to Eq. (13), Eq. (16), and triangle inequality, we can derive that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[||\boldsymbol{x}^{i,t+1}-\boldsymbol{x}^{t+1}||^2] \\
&=\frac{\eta_{\boldsymbol{x}}^2}{n}\sum_{i=1}^{n}\mathbb{E}[||\Delta_{\boldsymbol{x}}^{i,t}-\bar{g}_{\boldsymbol{x}}^{t}||^2] \\
&\leq\frac{\eta_{\boldsymbol{x}}^2}{n}\sum_{i=1}^{n}\mathbb{E}[||\Delta_{\boldsymbol{x}}^{i,t}||^2-||\bar{g}_{\boldsymbol{x}}^{t}||^2] \\
&\leq\frac{\eta_{\boldsymbol{x}}^2}{n}\sum_{i=1}^{n}\mathbb{E}[||\Delta_{\boldsymbol{x}}^{i,t}||^2].
\end{aligned} \tag{51}$$

Building upon the definition of $\Delta_{\boldsymbol{x}}^{i,t}$, and $\mathbb{E}\left[v_{\boldsymbol{x}}^{i,t}\right]=0$, we can obtain:

$$\begin{aligned}
&\mathbb{E}\left[||\Delta_{\boldsymbol{x}}^{i,t}||^2\right] \\
&=\mathbb{E}\left[\left\|\frac{1}{m}\sum_{k=1}^{m}\mathrm{clip}_{c_{\boldsymbol{x}}}(\nabla_{\boldsymbol{x}}L_{\lambda,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t}))+\frac{2c_{\boldsymbol{x}}}{m}v_{\boldsymbol{x}}^{i,t}\right\|^2\right] \\
&=\mathbb{E}\left[\left\|\frac{1}{m}\sum_{k=1}^{m}\mathrm{clip}_{c_{\boldsymbol{x}}}(\nabla_{\boldsymbol{x}}L_{\lambda,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t}))\right\|^2\right]+\mathbb{E}\left[\left\|\frac{2c_{\boldsymbol{x}}}{m}v_{\boldsymbol{x}}^{i,t}\right\|^2\right] \\
&\leq\mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m}||\mathrm{clip}_{c_{\boldsymbol{x}}}(\nabla_{\boldsymbol{x}}L_{\lambda,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t}))||^2\right]+\frac{4c_{\boldsymbol{x}}^2}{m^2}\sum_{k=1}^{m}\mathbb{E}\left[||v_{\boldsymbol{x}}^{i,t}||^2\right] \\
&\leq c_{\boldsymbol{x}}^2+\frac{4d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{m^2}.
\end{aligned} \tag{52}$$

Then, by combining Eq. (51) with Eq. (52), we have:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[||\boldsymbol{x}^{i,t+1}-\boldsymbol{x}^{t+1}||]\leq\frac{\eta_{\boldsymbol{x}}^2}{n}\sum_{i=1}^{n}\left(c_{\boldsymbol{x}}^2+\frac{4d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{m^2}\right)\leq\eta_{\boldsymbol{x}}^2\left(c_{\boldsymbol{x}}^2+\frac{4d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{m^2}\right). \tag{53}$$

Similarly, applying Eqs. (14) and (16) in conjunction with the triangle inequality, we can derive that:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[||\boldsymbol{y}^{i,t+1}-\boldsymbol{y}^{t+1}||^2] \\
&=\frac{\eta_{\boldsymbol{y}}^2}{n}\sum_{i=1}^{n}\mathbb{E}[||\Delta_{\boldsymbol{y}}^{i,t}-\bar{g}_{\boldsymbol{y}}^{t}||^2] \\
&\leq\frac{\eta_{\boldsymbol{y}}^2}{n}\sum_{i=1}^{n}\mathbb{E}[||\Delta_{\boldsymbol{y}}^{i,t}||^2] \\
&=\mathbb{E}\left[\left\|\frac{1}{m}\sum_{k=1}^{m}\mathrm{clip}_{c_{\boldsymbol{y}}}(\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t}))+\frac{2c_{\boldsymbol{y}}}{m}v_{\boldsymbol{y}}^{i,t}\right\|^2\right] \\
&\leq\mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m}||\mathrm{clip}_{c_{\boldsymbol{y}}}(\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t}))||^2\right]+\frac{4c_{\boldsymbol{y}}^2}{m^2}\sum_{k=1}^{m}\mathbb{E}\left[||v_{\boldsymbol{y}}^{i,t}||^2\right] \\
&\leq c_{\boldsymbol{y}}^2+\frac{4d_2c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{m^2}.
\end{aligned} \tag{54}$$

□

Lemma E.2 *Under Assumptions 4.1 and 4.2, we can obtain that the difference in the Lagrangian function between $(t+1)^{\text{th}}$ iteration and t^{th} iteration is upper bounded by:*

$$\begin{aligned}
&\mathbb{E}[L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^{t+1},\boldsymbol{z}^{t+1})]-\mathbb{E}[L_\lambda(\boldsymbol{x}^{t},\boldsymbol{y}^{t},\boldsymbol{z}^{t})]\\
&\leq -\frac{\eta_{\boldsymbol{x}}}{2}\mathbb{E}[||\nabla_{\boldsymbol{x}}L_\lambda(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)||^2]-\frac{\eta_{\boldsymbol{y}}}{2}\mathbb{E}[||\nabla_{\boldsymbol{y}}L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^t,\boldsymbol{z}^t)||^2]-\frac{\eta_{\boldsymbol{z}}}{2}\mathbb{E}[||\nabla_{\boldsymbol{z}}L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^{t+1},\boldsymbol{z}^t)||^2]\\
&+\frac{\eta_{\boldsymbol{x}}G_{\boldsymbol{x}}^2}{2c_{\boldsymbol{x}}}+\frac{L_{\boldsymbol{x}}\eta_{\boldsymbol{x}}^2}{2}\left({c_{\boldsymbol{x}}}^2+\frac{d_1{\sigma_{\boldsymbol{x}}}^2}{nm}\right)+\frac{\eta_{\boldsymbol{y}}G_{\boldsymbol{y}}^4}{c_{\boldsymbol{y}}^2}+\eta_{\boldsymbol{y}}L^2\left(\frac{1}{n}\sum_{i=1}^n\mathbb{E}[||\boldsymbol{x}^{i,t+1}-\boldsymbol{x}^{t+1}||^2]\right)\\
&+\frac{L\eta_{\boldsymbol{y}}^2}{2}\left(c_{\boldsymbol{y}}^2+\frac{d_2\sigma_{\boldsymbol{y}}^2}{nm}\right)+\frac{\eta_{\boldsymbol{z}}G_{\boldsymbol{z}}^4}{c_{\boldsymbol{z}}^2}+\frac{L\eta_{\boldsymbol{z}}^2}{2}\left(c_{\boldsymbol{z}}^2+\frac{d_3\sigma_{\boldsymbol{z}}^2}{nm}\right)\\
&+2\eta_{\boldsymbol{z}}L^2\left(\frac{1}{n}\sum_{i=1}^n\mathbb{E}[||\boldsymbol{x}^{i,t+1}-\boldsymbol{x}^{t+1}||^2]\right)+2\eta_{\boldsymbol{z}}L^2\left(\frac{1}{n}\sum_{i=1}^n\mathbb{E}[||\boldsymbol{y}^{i,t+1}-\boldsymbol{y}^{t+1}||^2]\right).
\end{aligned}\tag{55}$$

Proof of Lemma E.2:

According to the definition of $L_\lambda(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$ (i.e., $L_\lambda(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})=\frac{1}{n}\sum_i f_{1,i}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})+\lambda\sum_l o_l(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$) and Assumption 4.1, for any $\boldsymbol{x}_1,\boldsymbol{x}_2$, we can get:

$$\begin{aligned}
&||\nabla_{\boldsymbol{x}}L_\lambda(\boldsymbol{x}_1,\boldsymbol{y},\boldsymbol{z})-\nabla_{\boldsymbol{x}}L_\lambda(\boldsymbol{x}_2,\boldsymbol{y},\boldsymbol{z})||\\
&=\left|\left|\frac{1}{n}\sum_{i=1}^n(\nabla_{\boldsymbol{x}}f_{1,i}(\boldsymbol{x}_1,\boldsymbol{y},\boldsymbol{z})-\nabla_{\boldsymbol{x}}f_{1,i}(\boldsymbol{x}_2,\boldsymbol{y},\boldsymbol{z}))+2\lambda\sum\nolimits_l\text{diag}(\boldsymbol{a}_{x,l})(\boldsymbol{x}_1-\boldsymbol{x}_2)\right|\right|\\
&\leq\frac{1}{n}\sum_{i=1}^n||\nabla_{\boldsymbol{x}}f_{1,i}(\boldsymbol{x}_1,\boldsymbol{y},\boldsymbol{z})-\nabla_{\boldsymbol{x}}f_{1,i}(\boldsymbol{x}_2,\boldsymbol{y},\boldsymbol{z})||+2\lambda\left|\left|\text{diag}\left(\sum\nolimits_l\boldsymbol{a}_{x,l}\right)(\boldsymbol{x}_1-\boldsymbol{x}_2)\right|\right|\\
&\leq L||\boldsymbol{x}_1-\boldsymbol{x}_2||+2\lambda\left|\left|\sum\nolimits_l\boldsymbol{a}_{x,l}\right|\right|_\infty||\boldsymbol{x}_1-\boldsymbol{x}_2||\\
&\leq\hat{L}||\boldsymbol{x}_1-\boldsymbol{x}_2||.
\end{aligned}\tag{56}$$

where the first inequality follows from the triangle inequality, and $\hat{L}=L+2\lambda\max\{\|\sum_l\boldsymbol{a}_{x,l}\|_\infty,\|\sum_l\boldsymbol{a}_{y,l}\|_\infty,\|\sum_l\boldsymbol{a}_{z,l}\|_\infty\}$ is a positive constant. Consequently, by virtue of Eq. (56), we obtain:

$$\begin{aligned}
&\mathbb{E}[L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^t,\boldsymbol{z}^t)]\\
&\leq\mathbb{E}\left[L_\lambda(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)\right]+\mathbb{E}\left[\left\langle\nabla_{\boldsymbol{x}}L_\lambda(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t),\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\right\rangle\right]+\frac{\hat{L}}{2}\mathbb{E}\left[||\boldsymbol{x}^{t+1}-\boldsymbol{x}^t||^2\right]\\
&=L_\lambda(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)-\eta_{\boldsymbol{x}}\langle\nabla_{\boldsymbol{x}}L_\lambda(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t),\mathbb{E}[\bar{g}_{\boldsymbol{x}}^t]\rangle+\frac{\hat{L}\eta_{\boldsymbol{x}}^2}{2}\mathbb{E}\left[||\bar{g}_{\boldsymbol{x}}^t||^2\right].
\end{aligned}\tag{57}$$

Let $b_{\boldsymbol{x}}^t=\mathbb{E}[\bar{g}_{\boldsymbol{x}}^t]-\nabla_{\boldsymbol{x}}L_\lambda(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)$. By invoking the triangle inequality in conjunction with the fact that $\mathbb{E}\left[v_{\boldsymbol{x}}^{i,t}\right]=0$, we arrive at:

$$\begin{aligned}
&||b_{\boldsymbol{x}}^t||\\
&=\left|\left|\frac{1}{nm}\sum_{i=1}^n\sum_{k=1}^m\left(\mathbb{E}[\text{clip}_{c_{\boldsymbol{x}}}(\nabla_{\boldsymbol{x}}L_{\lambda,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t}))]-\nabla_{\boldsymbol{x}}L_{\lambda,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)\right)\right|\right|\\
&\leq\frac{1}{nm}\sum_{i=1}^n\sum_{k=1}^m\left\|\mathbb{E}[\text{clip}_{c_{\boldsymbol{x}}}(\nabla_{\boldsymbol{x}}L_{\lambda,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t}))]-\nabla_{\boldsymbol{x}}L_{\lambda,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)\right\|.
\end{aligned}\tag{58}$$

Additionally, based on the definition of $L_{\lambda,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t})$, we have that:

$$\begin{aligned}
&\mathbb{E}\left[||\nabla_{\boldsymbol{x}}L_{\lambda,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t})||^2\right]\\
&\leq 2\mathbb{E}\left[||\nabla_{\boldsymbol{x}}f_{1,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t})||^2\right]+2\mathbb{E}\left[||\nabla_{\boldsymbol{x}}\sum\nolimits_l o_l(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t})||^2\right]\\
&\leq 2G_f^2+2\left(2||\sum\nolimits_l\boldsymbol{a}_{x,l}||_\infty D_{\boldsymbol{x}}+||\sum\nolimits_l\boldsymbol{b}_{x,l}||\right)^2.
\end{aligned}\tag{59}$$

Let constant $G_{\boldsymbol{x}} \geq \sqrt{2G_f^2 + 2\left(2\left|\left|\sum_l \boldsymbol{a}_{x,l}\right|\right|_\infty D_{\boldsymbol{x}} + \left|\left|\sum_l \boldsymbol{b}_{x,l}\right|\right|\right)^2}$, and define the indicator function $\mathbb{I}_{||\nabla_{\boldsymbol{x}} L_{\lambda,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t})||>c_{\boldsymbol{x}}}$, by combining the Cauchy-Schwarz inequality with Markov's inequality, we can derive:

$$\begin{aligned}
&\left\|\mathbb{E}[\mathrm{clip}_{c_{\boldsymbol{x}}}(\nabla_{\boldsymbol{x}} L_{\lambda,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t}))] - \nabla_{\boldsymbol{x}} L_{\lambda,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)\right\| \\
&\leq \mathbb{E}\left[||\nabla_{\boldsymbol{x}} L_{\lambda,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t})|| \cdot \mathbb{I}_{||\nabla_{\boldsymbol{x}} L_{\lambda,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t})||>c_{\boldsymbol{x}}}\right] \\
&\leq \sqrt{\mathbb{E}\left[||\nabla_{\boldsymbol{x}} L_{\lambda,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t})||^2\right]}\sqrt{P(||\nabla_{\boldsymbol{x}} L_{\lambda,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t})|| > c_{\boldsymbol{x}})} \\
&\leq \sqrt{G_{\boldsymbol{x}}^2}\sqrt{\frac{\mathbb{E}\left[||\nabla_{\boldsymbol{x}} L_{\lambda,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t})||^2\right]}{c_{\boldsymbol{x}}^2}} \\
&\leq \frac{G_{\boldsymbol{x}}^2}{c_{\boldsymbol{x}}}.
\end{aligned} \tag{60}$$

According to Eq. (58) and Eq. (60), we can obtain:

$$||b_{\boldsymbol{x}}^t|| \leq \frac{1}{n}\sum_{i=1}^n \frac{1}{m}\sum_{k=1}^m \frac{G_{\boldsymbol{x}}^2}{c_{\boldsymbol{x}}} = \frac{G_{\boldsymbol{x}}^2}{c_{\boldsymbol{x}}}. \tag{61}$$

Based on Cauchy-Schwarz inequality and $\mathbb{E}[||\sum_{i=1}^n v_{\boldsymbol{x}}^{i,t}||^2] = \sum_{i=1}^n \mathbb{E}[||v_{\boldsymbol{x}}^{i,t}||^2]$, we have:

$$\begin{aligned}
&\mathbb{E}\left[||\bar{g}_{\boldsymbol{x}}^t||^2\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{nm}\sum_{i=1}^n\sum_{k=1}^m \mathrm{clip}_{c_{\boldsymbol{x}}}(\nabla_{\boldsymbol{x}} L_{\lambda,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t})) + \frac{1}{n}\sum_{i=1}^n \frac{2c_{\boldsymbol{x}}}{m} v_{\boldsymbol{x}}^{i,t}\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{nm}\sum_{i=1}^n\sum_{k=1}^m \mathrm{clip}_{c_{\boldsymbol{x}}}(\nabla_{\boldsymbol{x}} L_{\lambda,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t}))\right\|^2\right] + \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \frac{2c_{\boldsymbol{x}}}{m} v_{\boldsymbol{x}}^{i,t}\right\|^2\right] \\
&\leq \frac{1}{nm}\sum_{i=1}^n\sum_{k=1}^m \mathbb{E}\left\|\mathrm{clip}_{c_{\boldsymbol{x}}}(\nabla_{\boldsymbol{x}} L_{\lambda,i}(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t;\xi_{\boldsymbol{x}}^{i,k,t}))\right\|^2 + \frac{4c_{\boldsymbol{x}}^2}{n^2m^2}\sum_{i=1}^n \mathbb{E}[||v_{\boldsymbol{x}}^{i,t}||^2] \\
&= c_{\boldsymbol{x}}^2 + \frac{4d_1 c_{\boldsymbol{x}}^2 {\sigma_{\boldsymbol{x}}}^2}{nm^2}.
\end{aligned} \tag{62}$$

In addition, according to inequality $-\langle a,b\rangle \leq \frac{1}{2}||a||^2 + \frac{1}{2}||b||^2$, we can obtain:

$$-\eta_{\boldsymbol{x}}\mathbb{E}[\langle \nabla_{\boldsymbol{x}} L_\lambda(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t), b_{\boldsymbol{x}}^t\rangle] \leq \frac{\eta_{\boldsymbol{x}}}{2}\mathbb{E}[||\nabla_{\boldsymbol{x}} L_\lambda(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)||^2] + \frac{\eta_{\boldsymbol{x}}}{2}||b_{\boldsymbol{x}}^t||^2. \tag{63}$$

Combining Eqs. (57), (62), and (63), we obtain:

$$\begin{aligned}
&\mathbb{E}[L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^t,\boldsymbol{z}^t)] - \mathbb{E}[L_\lambda(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)] \\
&\leq -\frac{\eta_{\boldsymbol{x}}}{2}\mathbb{E}[||\nabla_{\boldsymbol{x}} L_\lambda(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)||^2] + \frac{\eta_{\boldsymbol{x}}}{2}||b_{\boldsymbol{x}}^t||^2 + \frac{\hat{L}\eta_{\boldsymbol{x}}^2}{2}\mathbb{E}[||\bar{g}_{\boldsymbol{x}}^t||^2] \\
&\leq -\frac{\eta_{\boldsymbol{x}}}{2}\mathbb{E}[||\nabla_{\boldsymbol{x}} L_\lambda(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)||^2] + \frac{\eta_{\boldsymbol{x}} G_{\boldsymbol{x}}^2}{2c_{\boldsymbol{x}}} + \frac{\hat{L}\eta_{\boldsymbol{x}}^2}{2}\left(c_{\boldsymbol{x}}^2 + \frac{4d_1 c_{\boldsymbol{x}}^2 {\sigma_{\boldsymbol{x}}}^2}{nm^2}\right).
\end{aligned} \tag{64}$$

Similarly, for the second-level variable, we first define:

$$b_{\boldsymbol{y}}^t = \mathbb{E}[\bar{g}_{\boldsymbol{y}}^t] - \frac{1}{n}\sum_{i=1}^n \nabla_{\boldsymbol{y}} L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^t,\boldsymbol{z}^t). \tag{65}$$

According to triangle inequality and $\mathbb{E}\left[v_{i,k,t}^{y}\right]=0$, we can derive that

$$\begin{aligned}
&||b_{\boldsymbol{y}}^{t}|| \\
&=\left\|\frac{1}{nm}\sum_{i=1}^{n}\sum_{k=1}^{m}\left(\mathbb{E}[\mathrm{clip}_{c_{\boldsymbol{y}}}(\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^{t},\boldsymbol{z}^{t};\xi_{\boldsymbol{y}}^{i,k,t}))]-\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^{t},\boldsymbol{z}^{t})\right)\right\| \\
&\leq\frac{1}{nm}\sum_{i=1}^{n}\sum_{k=1}^{m}||\mathbb{E}[\mathrm{clip}_{c_{\boldsymbol{y}}}(\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^{t},\boldsymbol{z}^{t};\xi_{\boldsymbol{y}}^{i,k,t}))]-\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^{t},\boldsymbol{z}^{t})||.
\end{aligned}\tag{66}$$

Based on the definition of $L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^{t},\boldsymbol{z}^{t};\xi_{\boldsymbol{y}}^{i,k,t})$, we have:

$$\begin{aligned}
&\mathbb{E}\left[||\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^{t},\boldsymbol{z}^{t};\xi_{\boldsymbol{y}}^{i,k,t})||^{2}\right] \\
&\leq 2\mathbb{E}\left[||\nabla_{\boldsymbol{y}}f_{1,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^{t},\boldsymbol{z}^{t};\xi_{\boldsymbol{y}}^{i,k,t})||^{2}\right]+2\mathbb{E}\left[||\nabla_{\boldsymbol{y}}\sum\nolimits_{l}o_{l}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^{t},\boldsymbol{z}^{t};\xi_{\boldsymbol{y}}^{i,k,t})||^{2}\right] \\
&\leq 2G_{f}^{2}+2\left(2||\sum\nolimits_{l}\boldsymbol{a}_{y,l}||_{\infty}D_{\boldsymbol{y}}+||\sum\nolimits_{l}\boldsymbol{b}_{y,l}||\right)^{2}.
\end{aligned}\tag{67}$$

Let $\mathbb{I}_{||\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^{t},\boldsymbol{z}^{t};\xi_{\boldsymbol{y}}^{i,k,t})||>c_{\boldsymbol{y}}}$ denote the indicator function. By leveraging the Cauchy-Schwarz and Markov's inequalities, and defining the constant $G_{\boldsymbol{y}}\geq\sqrt{2G_{f}^{2}+2\left(2\|\sum_{l}\boldsymbol{a}_{y,l}\|_{\infty}D_{\boldsymbol{y}}+\|\sum_{l}\boldsymbol{b}_{y,l}\|\right)^{2}}$, we obtain:

$$\begin{aligned}
&\left\|\mathbb{E}\left[\mathrm{clip}_{c_{\boldsymbol{y}}}(\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^{t},\boldsymbol{z}^{t};\xi_{\boldsymbol{y}}^{i,k,t}))\right]-\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^{t},\boldsymbol{z}^{t})\right\| \\
&\leq\mathbb{E}\left[||\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^{t},\boldsymbol{z}^{t};\xi_{\boldsymbol{y}}^{i,k,t})||\cdot\mathbb{I}_{||\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^{t},\boldsymbol{z}^{t};\xi_{\boldsymbol{y}}^{i,k,t})||>c_{\boldsymbol{y}}}\right] \\
&\leq\sqrt{\mathbb{E}[||\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^{t},\boldsymbol{z}^{t};\xi_{\boldsymbol{y}}^{i,k,t})||^{2}]}\sqrt{P(||\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^{t},\boldsymbol{z}^{t};\xi_{\boldsymbol{y}}^{i,k,t})||>c_{\boldsymbol{y}})} \\
&\leq\sqrt{G_{\boldsymbol{y}}^{2}}\sqrt{\frac{G_{\boldsymbol{y}}^{2}}{c_{\boldsymbol{y}}^{2}}} \\
&=\frac{G_{\boldsymbol{y}}^{2}}{c_{\boldsymbol{y}}}.
\end{aligned}\tag{68}$$

Thus, by combining Eq. (66) with Eq. (68), we have:

$$||b_{\boldsymbol{y}}^{t}||\leq\frac{1}{nm}\sum_{i=1}^{n}\sum_{k=1}^{m}\frac{G_{\boldsymbol{y}}^{2}}{c_{\boldsymbol{y}}}=\frac{G_{\boldsymbol{y}}^{2}}{c_{\boldsymbol{y}}}.\tag{69}$$

Additionally, we can also obtain:

$$\begin{aligned}
&\mathbb{E}[||\bar{g}_{\boldsymbol{y}}^{t}||^{2}] \\
&=\mathbb{E}\left[\left\|\frac{1}{nm}\sum_{i=1}^{n}\sum_{k=1}^{m}\mathrm{clip}_{c_{\boldsymbol{y}}}(\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^{t},\boldsymbol{z}^{t};\xi_{\boldsymbol{y}}^{i,k,t}))+\frac{1}{n}\sum_{i=1}^{n}\frac{2c_{\boldsymbol{y}}}{m}v_{\boldsymbol{y}}^{i,t}\right\|^{2}\right] \\
&=\mathbb{E}\left[\left\|\frac{1}{nm}\sum_{i=1}^{n}\sum_{k=1}^{m}\mathrm{clip}_{c_{\boldsymbol{y}}}(\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^{t},\boldsymbol{z}^{t};\xi_{\boldsymbol{y}}^{i,k,t}))\right\|^{2}\right]+\mathbb{E}\left[\left\|\frac{1}{n}\frac{2c_{\boldsymbol{y}}}{m}\sum_{i=1}^{n}v_{\boldsymbol{y}}^{i,t}\right\|^{2}\right] \\
&\leq c_{\boldsymbol{y}}^{2}+\frac{4d_{2}c_{\boldsymbol{y}}^{2}\sigma_{\boldsymbol{y}}^{2}}{nm^{2}}.
\end{aligned}\tag{70}$$

Similarly, according to the definition of $L_{\lambda}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$ and Assumption 4.1, for any $\boldsymbol{y}_{1},\boldsymbol{y}_{2}$, we can get:

$$\begin{aligned}
&||\nabla_{\boldsymbol{y}}L_{\lambda}(\boldsymbol{x},\boldsymbol{y}_{1},\boldsymbol{z})-\nabla_{\boldsymbol{y}}L_{\lambda}(\boldsymbol{x},\boldsymbol{y}_{2},\boldsymbol{z})|| \\
&=\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla_{\boldsymbol{y}}f_{1,i}(\boldsymbol{x},\boldsymbol{y}_{1},\boldsymbol{z})-\nabla_{\boldsymbol{y}}f_{1,i}(\boldsymbol{x},\boldsymbol{y}_{2},\boldsymbol{z})\right)+2\lambda\sum\nolimits_{l}\mathrm{diag}(\boldsymbol{a}_{y,l})(\boldsymbol{y}_{1}-\boldsymbol{y}_{2})\right\| \\
&\leq L||\boldsymbol{y}_{1}-\boldsymbol{y}_{2}||+2\lambda\left\|\sum\nolimits_{l}\boldsymbol{a}_{y,l}\right\|_{\infty}||\boldsymbol{y}_{1}-\boldsymbol{y}_{2}|| \\
&\leq\hat{L}||\boldsymbol{y}_{1}-\boldsymbol{y}_{2}||.
\end{aligned}\tag{71}$$

Subsequently, we can obtain:

$$\begin{aligned}
&\mathbb{E}[L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^{t+1},\boldsymbol{z}^t)]\\
&\leq \mathbb{E}[L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^{t},\boldsymbol{z}^t)]+\mathbb{E}\left[\langle\nabla_{\boldsymbol{y}}L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^{t},\boldsymbol{z}^t),\boldsymbol{y}^{t+1}-\boldsymbol{y}^t\rangle\right]+\frac{\hat{L}}{2}\mathbb{E}[||\boldsymbol{y}^{t+1}-\boldsymbol{y}^t||^2]\\
&\leq -\eta_{\boldsymbol{y}}\mathbb{E}\left[\langle\nabla_{\boldsymbol{y}}L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^{t},\boldsymbol{z}^t),\bar{g}^t_{\boldsymbol{y}}\rangle\right]+\frac{\hat{L}\eta^2_{\boldsymbol{y}}}{2}\mathbb{E}[||\bar{g}^t_{\boldsymbol{y}}||^2].
\end{aligned}\tag{72}$$

According to Eq. (65), we have:

$$\begin{aligned}
&\mathbb{E}[\bar{g}^t_{\boldsymbol{y}}]=b^t_{\boldsymbol{y}}+\tfrac{1}{n}\sum_{i=1}^n\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{t+1},\boldsymbol{y}^t,\boldsymbol{z}^t)\\
&\underbrace{+\frac{1}{n}\sum_{i=1}^n\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^t,\boldsymbol{z}^t)-\frac{1}{n}\sum_{i=1}^n\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{t+1},\boldsymbol{y}^t,\boldsymbol{z}^t)}_{e^t_{\boldsymbol{y}}},
\end{aligned}\tag{73}$$

it follows that

$$\begin{aligned}
&-\eta_{\boldsymbol{y}}\langle\nabla_{\boldsymbol{y}}L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^t,\boldsymbol{z}^t),\mathbb{E}[\bar{g}^t_{\boldsymbol{y}}]\rangle\\
&=-\eta_{\boldsymbol{y}}||\nabla_{\boldsymbol{y}}L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^t,\boldsymbol{z}^t)||^2-\eta_{\boldsymbol{y}}\langle\nabla_{\boldsymbol{y}}L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^t,\boldsymbol{z}^t),e^t_{\boldsymbol{y}}+b^t_{\boldsymbol{y}}\rangle.
\end{aligned}\tag{74}$$

According to Young's inequality, i.e., $-\langle a,b+c\rangle\leq\left(\frac{1}{4}||a||^2+||b||^2\right)+\left(\frac{1}{4}||a||^2+||c||^2\right)=\frac{1}{2}||a||^2+||b||^2+||c||^2$, we can derive that

$$\begin{aligned}
&-\eta_{\boldsymbol{y}}\langle\nabla_{\boldsymbol{y}}L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^t,\boldsymbol{z}^t),\mathbb{E}[\bar{g}^t_{\boldsymbol{y}}]\rangle\\
&\leq-\eta_{\boldsymbol{y}}||\nabla_{\boldsymbol{y}}L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^t,\boldsymbol{z}^t)||^2+\frac{\eta_{\boldsymbol{y}}}{2}||\nabla_{\boldsymbol{y}}L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^t,\boldsymbol{z}^t)||^2+\eta_{\boldsymbol{y}}||e^t_{\boldsymbol{y}}||^2+\eta_{\boldsymbol{y}}||b^t_{\boldsymbol{y}}||^2\\
&=-\frac{\eta_{\boldsymbol{y}}}{2}||\nabla_{\boldsymbol{y}}L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^t,\boldsymbol{z}^t)||^2+\eta_{\boldsymbol{y}}||e^t_{\boldsymbol{y}}||^2+\eta_{\boldsymbol{y}}||b^t_{\boldsymbol{y}}||^2.
\end{aligned}\tag{75}$$

For $||e^t_{\boldsymbol{y}}||^2$ in Eq. (75), according to Assumption 4.1, we have:

$$\begin{aligned}
&||e^t_{\boldsymbol{y}}||^2\\
&=||\tfrac{1}{n}\sum_{i=1}^n\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^t,\boldsymbol{z}^t)-\tfrac{1}{n}\sum_{i=1}^n\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{t+1},\boldsymbol{y}^t,\boldsymbol{z}^t)||^2\\
&\leq\tfrac{1}{n}\sum_{i=1}^n||\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^t,\boldsymbol{z}^t)-\nabla_{\boldsymbol{y}}L_{\lambda,i}(\boldsymbol{x}^{t+1},\boldsymbol{y}^t,\boldsymbol{z}^t)||^2\\
&\leq\hat{L}^2\left(\tfrac{1}{n}\sum_{i=1}^n||\boldsymbol{x}^{i,t+1}-\boldsymbol{x}^{t+1}||^2\right).
\end{aligned}\tag{76}$$

Consequently, by combining Eq. (69), Eq. (72), Eq. (75) with Eq. (76), we can derive that

$$\begin{aligned}
&\mathbb{E}[L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^{t+1},\boldsymbol{z}^t)]\\
&\leq\mathbb{E}[L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^{t},\boldsymbol{z}^t)]-\frac{\eta_{\boldsymbol{y}}}{2}\mathbb{E}[||\nabla_{\boldsymbol{y}}L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^t,\boldsymbol{z}^t)||^2]+\frac{\eta_{\boldsymbol{y}}G^4_{\boldsymbol{y}}}{c^2_{\boldsymbol{y}}}\\
&+\eta_{\boldsymbol{y}}\hat{L}^2\left(\frac{1}{n}\sum_{i=1}^n\mathbb{E}[||\boldsymbol{x}^{i,t+1}-\boldsymbol{x}^{t+1}||^2]\right)+\frac{\hat{L}\eta^2_{\boldsymbol{y}}}{2}\left(c^2_{\boldsymbol{y}}+\frac{4d_2c^2_{\boldsymbol{y}}\sigma^2_{\boldsymbol{y}}}{nm^2}\right).
\end{aligned}\tag{77}$$

For the third-level variable $\boldsymbol{z}$, we define $b^t_{\boldsymbol{z}}$ as:

$$b^t_{\boldsymbol{z}}=\mathbb{E}[\bar{g}^t_{\boldsymbol{z}}]-\frac{1}{n}\sum_{i=1}^n\nabla_{\boldsymbol{z}}L_{\lambda,i}(\boldsymbol{x}^{i,t+1},\boldsymbol{y}^{i,t+1},\boldsymbol{z}^t).\tag{78}$$

Likewise, building upon the definition of $L_{\lambda,i}(\boldsymbol{x}^{i,t+1}, \boldsymbol{y}^{i,t+1}, \boldsymbol{z}^t; \xi_{\boldsymbol{z}}^{i,k,t})$, we can get:

$$\begin{aligned}
&\mathbb{E}\left[||\nabla_{\boldsymbol{z}} L_{\lambda,i}(\boldsymbol{x}^{i,t+1}, \boldsymbol{y}^{i,t+1}, \boldsymbol{z}^t; \xi_{\boldsymbol{z}}^{i,k,t})||^2\right] \\
&\leq 2\mathbb{E}\left[||\nabla_{\boldsymbol{z}} f_{1,i}(\boldsymbol{x}^{i,t+1}, \boldsymbol{y}^{i,t+1}, \boldsymbol{z}^t; \xi_{\boldsymbol{z}}^{i,k,t})||^2\right] + 2\mathbb{E}\left[||\nabla_{\boldsymbol{z}} \sum\nolimits_l o_l(\boldsymbol{x}^{i,t+1}, \boldsymbol{y}^{i,t+1}, \boldsymbol{z}^t; \xi_{\boldsymbol{z}}^{i,k,t})||^2\right] \\
&\leq 2G_f^2 + 2\left(2||\sum\nolimits_l \boldsymbol{a}_{z,l}||_\infty D_{\boldsymbol{z}} + ||\sum\nolimits_l \boldsymbol{b}_{z,l}||\right)^2.
\end{aligned} \tag{79}$$

Let constant $G_{\boldsymbol{z}} \geq \sqrt{2G_f^2 + 2\left(2||\sum_l \boldsymbol{a}_{z,l}||_\infty D_{\boldsymbol{z}} + ||\sum_l \boldsymbol{b}_{z,l}||\right)^2}$, based on indicator function $\mathbb{I}_{||\nabla_{\boldsymbol{z}} L_{\lambda,i}(\boldsymbol{x}^{i,t+1}, \boldsymbol{y}^{i,t+1}, \boldsymbol{z}^t; \xi_{\boldsymbol{z}}^{i,k,t})||>c_{\boldsymbol{z}}}$, Cauchy-Schwarz inequality and Markov's inequality, we have:

$$\begin{aligned}
&||b_{\boldsymbol{z}}^t|| \\
&= \left\| \frac{1}{nm} \sum_{i=1}^{n} \sum_{k=1}^{m} \left(\mathbb{E}\left[\mathrm{clip}_{c_{\boldsymbol{z}}}(\nabla_{\boldsymbol{z}} L_{\lambda,i}(t+1))\right] - \nabla_{\boldsymbol{z}} L_{\lambda,i}(\boldsymbol{x}^{i,t+1}, \boldsymbol{y}^{i,t+1}, \boldsymbol{z}^t)\right) \right\| \\
&\leq \frac{1}{nm} \sum_{i=1}^{n} \sum_{k=1}^{m} \mathbb{E}\left[||\nabla_{\boldsymbol{z}} L_{\lambda,i}(t+1)|| \cdot \mathbb{I}_{||\nabla_{\boldsymbol{z}} L_{\lambda,i}(t+1)||>c_{\boldsymbol{z}}}\right] \\
&\leq \frac{1}{nm} \sum_{i=1}^{n} \sum_{k=1}^{m} \sqrt{\mathbb{E}[||\nabla_{\boldsymbol{z}} L_{\lambda,i}(t+1)||^2]} \cdot \sqrt{P(||\nabla_{\boldsymbol{z}} L_{\lambda,i}(t+1)|| > c_{\boldsymbol{z}})} \\
&\leq \frac{1}{nm} \sum_{i=1}^{n} \sum_{k=1}^{m} \sqrt{\mathbb{E}[||\nabla_{\boldsymbol{z}} L_{\lambda,i}(t+1)||^2]} \cdot \sqrt{\frac{\mathbb{E}[||\nabla_{\boldsymbol{z}} L_{\lambda,i}(t+1)||^2]}{c_{\boldsymbol{z}}^2}} \\
&\leq \frac{1}{nm} \sum_{i=1}^{n} \sum_{k=1}^{m} \sqrt{G_{\boldsymbol{z}}^2} \cdot \sqrt{\frac{G_{\boldsymbol{z}}^2}{c_{\boldsymbol{z}}^2}} \\
&= \frac{G_{\boldsymbol{z}}^2}{c_{\boldsymbol{z}}},
\end{aligned} \tag{80}$$

where $\nabla_{\boldsymbol{z}} L_{\lambda,i}(t+1) = \nabla_{\boldsymbol{z}} L_{\lambda,i}(\boldsymbol{x}^{i,t+1}, \boldsymbol{y}^{i,t+1}, \boldsymbol{z}^t; \xi_{\boldsymbol{z}}^{i,k,t})$. In addition, we have:

$$\begin{aligned}
&\mathbb{E}[||\bar{g}_{\boldsymbol{z}}^t||^2] \\
&= \mathbb{E}\left[\left\| \frac{1}{nm} \sum_{i=1}^{n} \sum_{k=1}^{m} \mathrm{clip}_{c_{\boldsymbol{z}}}(\nabla_{\boldsymbol{z}} L_{\lambda,i}(\boldsymbol{x}^{i,t+1}, \boldsymbol{y}^{i,t+1}, \boldsymbol{z}^t; \xi_{\boldsymbol{z}}^{i,k,t})) \right\|^2\right] + \mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} \frac{2c_{\boldsymbol{z}}}{m} v_{\boldsymbol{z}}^{i,t} \right\|^2\right] \\
&\leq \mathbb{E}\left[\frac{1}{nm} \sum_{i=1}^{n} \sum_{k=1}^{m} \left\|\mathrm{clip}_{c_{\boldsymbol{z}}}(\nabla_{\boldsymbol{z}} L_{\lambda,i}(\boldsymbol{x}^{i,t+1}, \boldsymbol{y}^{i,t+1}, \boldsymbol{z}^t; \xi_{\boldsymbol{z}}^{i,k,t}))\right\|^2\right] + \frac{4c_{\boldsymbol{z}}^2}{n^2 m^2} \mathbb{E}\left[\left\| \sum_{i=1}^{n} v_{\boldsymbol{z}}^{i,t} \right\|^2\right] \\
&\leq c_{\boldsymbol{z}}^2 + \frac{4d_3 c_{\boldsymbol{z}}^2 \sigma_{\boldsymbol{z}}^2}{nm^2}.
\end{aligned} \tag{81}$$

According to Assumption 4.1, for any $\boldsymbol{z}_1, \boldsymbol{z}_2$, we have :

$$\begin{aligned}
&||\nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}_1) - \nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}_2)|| \\
&\leq L||\boldsymbol{z}_1 - \boldsymbol{z}_2|| + 2\lambda \left|\left|\sum\nolimits_l \boldsymbol{a}_{z,l}\right|\right|_\infty ||\boldsymbol{z}_1 - \boldsymbol{z}_2|| \\
&\leq \hat{L}||\boldsymbol{z}_1 - \boldsymbol{z}_2||.
\end{aligned} \tag{82}$$

Then, the following inequality is obtained:

$$\begin{aligned}
&\mathbb{E}[L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1})] - \mathbb{E}[L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t)] \\
&\leq \mathbb{E}\left[\langle \nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t), \boldsymbol{z}^{t+1} - \boldsymbol{z}^t \rangle\right] + \frac{\hat{L}}{2} \mathbb{E}[||\boldsymbol{z}^{t+1} - \boldsymbol{z}^t||^2] \\
&= -\eta_{\boldsymbol{z}} \mathbb{E}\left[\langle \nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t), \bar{g}_{\boldsymbol{z}}^t \rangle\right] + \frac{\hat{L}\eta_{\boldsymbol{z}}^2}{2} \mathbb{E}[||\bar{g}_{\boldsymbol{z}}^t||^2].
\end{aligned} \tag{83}$$

In addition, it follows from Eq. (78) that:

$$\begin{aligned}
&\mathbb{E}[\bar{g}_{\boldsymbol{z}}^t] \\
&= \tfrac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{z}} L_{\lambda,i}(\boldsymbol{x}^{i,t+1}, \boldsymbol{y}^{i,t+1}, \boldsymbol{z}^t) + b_{\boldsymbol{z}}^t \\
&= \nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t) + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\nabla_{\boldsymbol{z}} L_{\lambda,i}(\boldsymbol{x}^{i,t+1}, \boldsymbol{y}^{i,t+1}, \boldsymbol{z}^t) - \nabla_{\boldsymbol{z}} L_{\lambda,i}(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t)\right)}_{e_{\boldsymbol{z}}^t} + b_{\boldsymbol{z}}^t.
\end{aligned} \tag{84}$$

According to Young's inequality $-\langle a, b+c\rangle \leq \frac{1}{2}||a||^2 + ||b||^2 + ||c||^2$, we obtain:

$$\begin{aligned}
&-\eta_{\boldsymbol{z}}\langle \nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t), \nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t) + e_{\boldsymbol{z}}^t + b_{\boldsymbol{z}}^t\rangle \\
&\leq -\frac{\eta_{\boldsymbol{z}}}{2}||\nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t)||^2 + \eta_{\boldsymbol{z}}||e_{\boldsymbol{z}}^t||^2 + \eta_{\boldsymbol{z}}||b_{\boldsymbol{z}}^t||^2.
\end{aligned} \tag{85}$$

Moreover, by combining Cauchy-Schwarz inequality with Assumption 4.1, we have:

$$\begin{aligned}
\mathbb{E}[||e_{\boldsymbol{z}}^t||^2] &= \left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla_{\boldsymbol{z}} L_{\lambda,i}(\boldsymbol{x}^{i,t+1}, \boldsymbol{y}^{i,t+1}, \boldsymbol{z}^t) - \nabla_{\boldsymbol{z}} L_{\lambda,i}(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t)\right)\right\|^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla_{\boldsymbol{z}} L_{\lambda,i}(\boldsymbol{x}^{i,t+1}, \boldsymbol{y}^{i,t+1}, \boldsymbol{z}^t) - \nabla_{\boldsymbol{z}} L_{\lambda,i}(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t)\right\|^2 \\
&\leq \frac{2}{n}\sum_{i=1}^{n}\left\|\nabla_{\boldsymbol{z}} L_{\lambda,i}(\boldsymbol{x}^{i,t+1}, \boldsymbol{y}^{i,t+1}, \boldsymbol{z}^t) - \nabla_{\boldsymbol{z}} L_{\lambda,i}(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{i,t+1}, \boldsymbol{z}^t)\right\|^2 \\
&+ \frac{2}{n}\sum_{i=1}^{n}\left\|\nabla_{\boldsymbol{z}} L_{\lambda,i}(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{i,t+1}, \boldsymbol{z}^t) - \nabla_{\boldsymbol{z}} L_{\lambda,i}(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t)\right\|^2 \\
&\leq 2\hat{L}^2\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[||\boldsymbol{x}^{i,t+1} - \boldsymbol{x}^{t+1}||^2]\right) + 2\hat{L}^2\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[||\boldsymbol{y}^{i,t+1} - \boldsymbol{y}^{t+1}||^2]\right).
\end{aligned} \tag{86}$$

Thus, we can obtain that

$$\begin{aligned}
&\mathbb{E}[L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1})] \\
&\leq \mathbb{E}[L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t)] - \frac{\eta_{\boldsymbol{z}}}{2}\mathbb{E}[||\nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t)||^2] + \frac{\eta_{\boldsymbol{z}} G_{\boldsymbol{z}}^4}{c_{\boldsymbol{z}}^2} + \frac{\hat{L}\eta_{\boldsymbol{z}}^2}{2}\left(c_{\boldsymbol{z}}^2 + \frac{4d_3 c_{\boldsymbol{z}}^2 \sigma_{\boldsymbol{z}}^2}{nm^2}\right) \\
&+ 2\eta_{\boldsymbol{z}}\hat{L}^2\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[||\boldsymbol{x}^{i,t+1} - \boldsymbol{x}^{t+1}||^2]\right) + 2\eta_{\boldsymbol{z}}\hat{L}^2\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[||\boldsymbol{y}^{i,t+1} - \boldsymbol{y}^{t+1}||^2]\right).
\end{aligned} \tag{87}$$

Consequently, by combining Eq. (64), Eq. (77) with Eq. (87), we can derive:

$$\begin{aligned}
&\mathbb{E}[L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1})] - \mathbb{E}[L_\lambda(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)] \\
&\leq -\frac{\eta_{\boldsymbol{x}}}{2}\mathbb{E}[||\nabla_{\boldsymbol{x}} L_\lambda(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2] - \frac{\eta_{\boldsymbol{y}}}{2}\mathbb{E}[||\nabla_{\boldsymbol{y}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2] - \frac{\eta_{\boldsymbol{z}}}{2}\mathbb{E}[||\nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t)||^2] \\
&+ \frac{\eta_{\boldsymbol{x}} G_{\boldsymbol{x}}^2}{2c_{\boldsymbol{x}}} + \frac{\hat{L}\eta_{\boldsymbol{x}}^2}{2}\left({c_{\boldsymbol{x}}}^2 + \frac{4d_1 c_{\boldsymbol{x}}^2 {\sigma_{\boldsymbol{x}}}^2}{nm^2}\right) + \frac{\eta_{\boldsymbol{y}} G_{\boldsymbol{y}}^4}{c_{\boldsymbol{y}}^2} + \frac{\hat{L}\eta_{\boldsymbol{y}}^2}{2}\left(c_{\boldsymbol{y}}^2 + \frac{d_2 c_{\boldsymbol{y}}^2 \sigma_{\boldsymbol{y}}^2}{nm^2}\right) \\
&+ \eta_{\boldsymbol{y}}\hat{L}^2\left(\frac{1}{n}\sum_{i=1}^{n}E||\boldsymbol{x}^{i,t+1} - \boldsymbol{x}^{t+1}||^2\right) + \frac{\eta_{\boldsymbol{z}} G_{\boldsymbol{z}}^4}{c_{\boldsymbol{z}}^2} + \frac{\hat{L}\eta_{\boldsymbol{z}}^2}{2}\left(c_{\boldsymbol{z}}^2 + \frac{4d_3 c_{\boldsymbol{z}}^2 \sigma_{\boldsymbol{z}}^2}{nm^2}\right) \\
&+ 2\eta_{\boldsymbol{z}}\hat{L}^2\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[||\boldsymbol{x}^{i,t+1} - \boldsymbol{x}^{t+1}||^2]\right) + 2\eta_{\boldsymbol{z}}\hat{L}^2\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[||\boldsymbol{y}^{i,t+1} - \boldsymbol{y}^{t+1}||^2]\right).
\end{aligned} \tag{88}$$

□

Lemma E.3 *In $(t+1)^{\text{th}}$ iteration, we can bound the negative expected squared norms of the partial gradients of the Lagrangian function with respect to $\boldsymbol{y}$ and $\boldsymbol{z}$ as follows:*

$$\begin{aligned}
&-\mathbb{E}[||\nabla_{\boldsymbol{y}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2] \\
&\leq -\frac{1}{2}\mathbb{E}[||\nabla_{\boldsymbol{y}} L_\lambda(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2] + \hat{L}^2\eta_{\boldsymbol{x}}^2\left(c_{\boldsymbol{x}}^2 + \frac{4d_1 c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{nm^2}\right),
\end{aligned} \tag{89}$$

$$\begin{aligned}
&-\mathbb{E}[||\nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t)||^2] \\
&\leq -\frac{1}{2}\mathbb{E}[||\nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2] + 2\hat{L}^2\eta_{\boldsymbol{y}}^2\left(c_{\boldsymbol{y}}^2 + \frac{4d_2 c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{nm^2}\right) + 2\hat{L}^2\eta_{\boldsymbol{x}}^2\left(c_{\boldsymbol{x}}^2 + \frac{4d_1 c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{nm^2}\right).
\end{aligned} \tag{90}$$

Proof of Lemma E.3:

According to Eq. (16) and Eq. (62), we have:

$$\mathbb{E}[||\boldsymbol{x}^{t+1} - \boldsymbol{x}^t||^2] = \eta_{\boldsymbol{x}}^2\mathbb{E}[||\bar{g}_{\boldsymbol{x}}^t||^2] \leq \eta_{\boldsymbol{x}}^2\left(c_{\boldsymbol{x}}^2 + \frac{4d_1 c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{nm^2}\right). \tag{91}$$

Based on the inequality $-||a||^2 \leq -\frac{1}{2}||b||^2 + ||a-b||^2$, and let $a = \nabla_{\boldsymbol{y}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t, \boldsymbol{z}^t), b = \nabla_{\boldsymbol{y}} L_\lambda(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)$, we have

$$\begin{aligned}
&-\mathbb{E}[||\nabla_{\boldsymbol{y}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2] \\
&\leq -\frac{1}{2}\mathbb{E}[||\nabla_{\boldsymbol{y}} L_\lambda(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2] + \mathbb{E}[||\nabla_{\boldsymbol{y}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t, \boldsymbol{z}^t) - \nabla_{\boldsymbol{y}} L_\lambda(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2].
\end{aligned} \tag{92}$$

In addition, according to $\hat{L}$-smoothness in Eq. (56):

$$\mathbb{E}[||\nabla_{\boldsymbol{y}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t, \boldsymbol{z}^t) - \nabla_{\boldsymbol{y}} L_\lambda(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2] \leq \hat{L}^2\mathbb{E}[||\boldsymbol{x}^{t+1} - \boldsymbol{x}^t||^2]. \tag{93}$$

Consequently, by combining Eq. (91), Eq. (92) with Eq. (93), we can derive:

$$\begin{aligned}
&-\mathbb{E}[||\nabla_{\boldsymbol{y}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2] \\
&\leq -\frac{1}{2}\mathbb{E}[||\nabla_{\boldsymbol{y}} L_\lambda(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2] + \hat{L}^2\eta_{\boldsymbol{x}}^2\left(c_{\boldsymbol{x}}^2 + \frac{4d_1 c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{nm^2}\right).
\end{aligned} \tag{94}$$

Likewise, by combining Eq. (14) with Eq. (70), it follows that

$$\mathbb{E}[||\boldsymbol{y}^{t+1} - \boldsymbol{y}^t||^2] = \eta_{\boldsymbol{y}}^2\mathbb{E}[||\bar{g}_{\boldsymbol{y}}^t||^2] \leq \eta_{\boldsymbol{y}}^2\left(c_{\boldsymbol{y}}^2 + \frac{4d_2 c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{nm^2}\right). \tag{95}$$

According to the inequality $-||a||^2 \leq -\frac{1}{2}||b||^2 + ||a-b||^2$ and let $a = \nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t), b = \nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)$, we obtain:

$$\begin{aligned}
&-\mathbb{E}[||\nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t)||^2] \\
&\leq -\frac{1}{2}\mathbb{E}[||\nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2] + \mathbb{E}[||\nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t) - \nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2].
\end{aligned} \tag{96}$$

In addition, we have the following inequality based on the Cauchy-Schwarz inequality and $\hat{L}$-smoothness in Eq. (56):

$$\begin{aligned}
&\mathbb{E}[||\nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t) - \nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2] \\
&= \mathbb{E}[||(\nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t) - \nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t, \boldsymbol{z}^t)) \\
&+(\nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t, \boldsymbol{z}^t) - \nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t))||^2] \\
&\leq 2\mathbb{E}[||\nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^t) - \nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2] \\
&+2\mathbb{E}[||\nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t, \boldsymbol{z}^t) - \nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t)||^2] \\
&\leq 2\hat{L}^2\mathbb{E}[||\boldsymbol{y}^{t+1} - \boldsymbol{y}^t||^2] + 2\hat{L}^2\mathbb{E}[||\boldsymbol{x}^{t+1} - \boldsymbol{x}^t||^2].
\end{aligned} \tag{97}$$

Consequently, the following inequality is derived:

$$
\begin{aligned}
&-\mathbb{E}[||\nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^{t+1},\boldsymbol{z}^t)||^2] \\
&\leq -\frac{1}{2}\mathbb{E}[||\nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t},\boldsymbol{y}^{t},\boldsymbol{z}^t)||^2] + 2L^2\eta_{\boldsymbol{y}}^2\left(c_{\boldsymbol{y}}^2+\frac{4d_2c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{nm^2}\right)+2L^2\eta_{\boldsymbol{x}}^2\left(c_{\boldsymbol{x}}^2+\frac{4d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{nm^2}\right).
\end{aligned}
\tag{98}
$$

□

Subsequently, we use the obtained Lemmas to derive the Theorem 4.4. By combining Lemma E.2 with Lemma E.3, we can derive:

$$
\begin{aligned}
&\mathbb{E}[L_\lambda(\boldsymbol{x}^{t+1},\boldsymbol{y}^{t+1},\boldsymbol{z}^{t+1})]-\mathbb{E}[L_\lambda(\boldsymbol{x}^{t},\boldsymbol{y}^{t},\boldsymbol{z}^{t})] \\
&\leq -\frac{\eta_{\boldsymbol{x}}}{4}\mathbb{E}[||\nabla_{\boldsymbol{x}} L_\lambda(\boldsymbol{x}^{t},\boldsymbol{y}^{t},\boldsymbol{z}^t)||^2]-\frac{\eta_{\boldsymbol{y}}}{4}\mathbb{E}[||\nabla_{\boldsymbol{y}} L_\lambda(\boldsymbol{x}^{t},\boldsymbol{y}^{t},\boldsymbol{z}^t)||^2]-\frac{\eta_{\boldsymbol{z}}}{4}\mathbb{E}[||\nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t},\boldsymbol{y}^{t},\boldsymbol{z}^t)||^2] \\
&+\frac{\eta_{\boldsymbol{y}}\hat{L}^2\eta_{\boldsymbol{x}}^2}{2}\left(c_{\boldsymbol{x}}^2+\frac{4d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{nm^2}\right)+\hat{L}^2\eta_{\boldsymbol{z}}\eta_{\boldsymbol{y}}^2\left(c_{\boldsymbol{y}}^2+\frac{4d_2c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{nm^2}\right)+\hat{L}^2\eta_{\boldsymbol{z}}\eta_{\boldsymbol{x}}^2\left(c_{\boldsymbol{x}}^2+\frac{4d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{nm^2}\right) \\
&+\frac{\eta_{\boldsymbol{x}}G_{\boldsymbol{x}}^2}{2c_{\boldsymbol{x}}}+\frac{\hat{L}\eta_{\boldsymbol{x}}^2}{2}\left({c_{\boldsymbol{x}}}^2+\frac{4d_1c_{\boldsymbol{x}}^2{\sigma_{\boldsymbol{x}}}^2}{nm^2}\right)+\frac{\eta_{\boldsymbol{y}}G_{\boldsymbol{y}}^4}{c_{\boldsymbol{y}}^2}+\hat{L}^2\eta_{\boldsymbol{y}}\eta_{\boldsymbol{x}}^2\left(c_{\boldsymbol{x}}^2+\frac{4d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{m^2}\right) \\
&+\frac{\hat{L}\eta_{\boldsymbol{y}}^2}{2}\left(c_{\boldsymbol{y}}^2+\frac{4d_2c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{nm^2}\right)+\frac{\eta_{\boldsymbol{z}}G_{\boldsymbol{z}}^4}{c_{\boldsymbol{z}}^2}+\frac{\hat{L}\eta_{\boldsymbol{z}}^2}{2}\left(c_{\boldsymbol{z}}^2+\frac{4d_3c_{\boldsymbol{z}}^2\sigma_{\boldsymbol{z}}^2}{nm^2}\right) \\
&+2\hat{L}^2\eta_{\boldsymbol{z}}\eta_{\boldsymbol{x}}^2\left(c_{\boldsymbol{x}}^2+\frac{4d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{m^2}\right)+2\hat{L}^2\eta_{\boldsymbol{z}}\eta_{\boldsymbol{y}}^2\left(c_{\boldsymbol{y}}^2+\frac{4d_2c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{m^2}\right).
\end{aligned}
\tag{99}
$$

According to the step-size setting, let $\eta_{\boldsymbol{x}}=\eta_{\boldsymbol{y}}=\eta_{\boldsymbol{z}}=\eta$. Summing Eq. (99) over $t=0,\ldots,T-1$, we obtain

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\frac{\eta}{4}\left(\mathbb{E}[||\nabla_{\boldsymbol{x}} L_\lambda(\boldsymbol{x}^{t},\boldsymbol{y}^{t},\boldsymbol{z}^t)||^2]+\mathbb{E}[||\nabla_{\boldsymbol{y}} L_\lambda(\boldsymbol{x}^{t},\boldsymbol{y}^{t},\boldsymbol{z}^t)||^2]+\mathbb{E}[||\nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^{t},\boldsymbol{y}^{t},\boldsymbol{z}^t)||^2]\right) \\
&\leq L_\lambda(\boldsymbol{x}_0,\boldsymbol{y}_0,\boldsymbol{z}_0)-f^*+\frac{\hat{L}^2\eta^3T}{2}\left(c_{\boldsymbol{x}}^2+\frac{4d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{nm^2}\right)+\hat{L}^2\eta^3T\left(c_{\boldsymbol{y}}^2+\frac{4d_2c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{nm^2}\right) \\
&+\hat{L}^2\eta^3T\left(c_{\boldsymbol{x}}^2+\frac{4d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{nm^2}\right)+\frac{\eta G_{\boldsymbol{x}}^2T}{2c_{\boldsymbol{x}}}+\frac{\hat{L}\eta^2T}{2}\left(c_{\boldsymbol{x}}^2+\frac{4d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{nm^2}\right)+\frac{\eta G_{\boldsymbol{y}}^4T}{c_{\boldsymbol{y}}^2} \\
&+\hat{L}^2\eta^3T\left(c_{\boldsymbol{x}}^2+\frac{4d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{nm^2}\right)+\frac{\hat{L}\eta^2T}{2}\left(c_{\boldsymbol{y}}^2+\frac{4d_2c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{nm^2}\right)+\frac{\eta G_{\boldsymbol{z}}^4T}{c_{\boldsymbol{z}}^2} \\
&+\frac{\hat{L}\eta^2T}{2}\left(c_{\boldsymbol{z}}^2+\frac{4d_3c_{\boldsymbol{z}}^2\sigma_{\boldsymbol{z}}^2}{nm^2}\right)+2\hat{L}^2\eta^3T\left(c_{\boldsymbol{x}}^2+\frac{4d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{m^2}\right)+2\hat{L}^2\eta^3T\left(c_{\boldsymbol{y}}^2+\frac{4d_2c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{m^2}\right) \\
&=L_\lambda(\boldsymbol{x}_0,\boldsymbol{y}_0,\boldsymbol{z}_0)-f^*+\eta T\left(\frac{G_{\boldsymbol{x}}^4}{2c_{\boldsymbol{x}}^2}+\frac{G_{\boldsymbol{y}}^4}{c_{\boldsymbol{y}}^2}+\frac{G_{\boldsymbol{z}}^4}{c_{\boldsymbol{z}}^2}\right) \\
&+\frac{\hat{L}^2\eta^2T}{2}\left(c_{\boldsymbol{x}}^2+\frac{4d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{nm^2}+c_{\boldsymbol{y}}^2+\frac{4d_2c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{nm^2}+c_{\boldsymbol{z}}^2+\frac{4d_3c_{\boldsymbol{z}}^2\sigma_{\boldsymbol{z}}^2}{nm^2}\right) \\
&+\hat{L}^2\eta^3T\left(\frac{3}{2}c_{\boldsymbol{x}}^2+6\frac{d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{nm^2}+c_{\boldsymbol{y}}^2+\frac{4d_2c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{nm^2}+3c_{\boldsymbol{x}}^2+\frac{12d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{m^2}+2c_{\boldsymbol{y}}^2+\frac{8d_2c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{m^2}\right).
\end{aligned}
\tag{100}
$$

Thus, we have that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}[||\nabla_{\boldsymbol{x}} L_\lambda(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)||^2]+\mathbb{E}[||\nabla_{\boldsymbol{y}} L_\lambda(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)||^2]+\mathbb{E}[||\nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)||^2\right)\\
&\leq \frac{(4L_\lambda(\boldsymbol{x}_0,\boldsymbol{y}_0,\boldsymbol{z}_0)-f^*)}{\eta T}+\left(\frac{2G_{\boldsymbol{x}}^4}{c_{\boldsymbol{x}}^2}+\frac{4G_{\boldsymbol{y}}^4}{c_{\boldsymbol{y}}^2}+\frac{4G_{\boldsymbol{z}}^4}{c_{\boldsymbol{z}}^2}\right)\\
&+2\hat{L}^2\eta\left(c_{\boldsymbol{x}}^2+\frac{4d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{nm^2}+c_{\boldsymbol{y}}^2+\frac{4d_2c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{nm^2}+c_{\boldsymbol{z}}^2+\frac{4d_3c_{\boldsymbol{z}}^2\sigma_{\boldsymbol{z}}^2}{nm^2}\right)\\
&+\hat{L}^2\eta^2\left(6\left(c_{\boldsymbol{x}}^2+\frac{4d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{nm^2}\right)+4\left(c_{\boldsymbol{y}}^2+\frac{4d_2c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{nm^2}\right)+12\left(c_{\boldsymbol{x}}^2+\frac{4d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{m^2}\right)+8\left(c_{\boldsymbol{y}}^2+\frac{4d_2c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{m^2}\right)\right).
\end{aligned}\tag{101}$$

Under the step-size setting $\eta_{\boldsymbol{x}}=\eta_{\boldsymbol{y}}=\eta_{\boldsymbol{z}}=\eta=\frac{1}{\sqrt{T}}$, we can derive the following bound:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}||\nabla L_\lambda(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)||^2\\
&=\frac{1}{T}\sum_{t=0}^{T-1}\left(||\nabla_{\boldsymbol{x}} L_\lambda(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)||^2+||\nabla_{\boldsymbol{y}} L_\lambda(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)||^2+||\nabla_{\boldsymbol{z}} L_\lambda(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)||^2\right)\\
&\leq \frac{(4L_\lambda(\boldsymbol{x}_0,\boldsymbol{y}_0,\boldsymbol{z}_0)-f^*)}{\sqrt{T}}+\left(\frac{2G_{\boldsymbol{x}}^4}{c_{\boldsymbol{x}}^2}+\frac{4G_{\boldsymbol{y}}^4}{c_{\boldsymbol{y}}^2}+\frac{4G_{\boldsymbol{z}}^4}{c_{\boldsymbol{z}}^2}\right)\\
&+\frac{2\hat{L}^2(c_{\boldsymbol{x}}^2+c_{\boldsymbol{y}}^2+c_{\boldsymbol{z}}^2)}{\sqrt{T}}+\frac{8\hat{L}^2}{\sqrt{T}}\left(\frac{d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{nm^2}+\frac{d_2c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{nm^2}+\frac{d_3c_{\boldsymbol{z}}^2\sigma_{\boldsymbol{z}}^2}{nm^2}\right)\\
&+\frac{1}{T}\hat{L}^2\left(18c_{\boldsymbol{x}}^2+12c_{\boldsymbol{y}}^2+\frac{24d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{nm^2}+\frac{16d_2c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{nm^2}+\frac{48d_1c_{\boldsymbol{x}}^2\sigma_{\boldsymbol{x}}^2}{m^2}+\frac{32d_2c_{\boldsymbol{y}}^2\sigma_{\boldsymbol{y}}^2}{m^2}\right).
\end{aligned}\tag{102}$$

□

Remark E.4 *In the context of differentially private optimization, the clipping operation inevitably incurs an additional bias [10, 115, 75]. The theoretical results established in Eq. (102) can be decomposed into two primary terms: an optimization bound and a clipping bias. Notably, the optimization term exhibits a convergence rate of $\mathcal{O}(1/\sqrt{T})$, matching the performance of state-of-the-art distributed trilevel optimization algorithms [44, 42]. Furthermore, the clipping bias term is shown to be tight, and the convergence neighborhood shrinks as clipping thresholds $c_{\boldsymbol{x}}, c_{\boldsymbol{y}}, c_{\boldsymbol{z}}$ increased, consistent with analyses in [48]. In addition, as indicated by the last two terms in Eq. (102), the proposed DPDTO achieves a speedup with respect to the number of workers n. This indicates that the proposed algorithm scales efficiently, making it suitable for large-scale distributed systems.*

F Proof of Theorem 4.5

Theorem 4.5 (Privacy of DPDTO) There exist constants β_0, β_1 for any privacy budget $\epsilon \leq 2\beta_0\beta_1\left(3q_1^2T+4q_2^2\left\lfloor\frac{T}{K}\right\rfloor\right)$ with sampling rates $q_1=m/M$ and $q_2=\tilde{m}/M$, where M represents the total number of data samples, the proposed algorithm is (ϵ,δ)-differentially private for any $\delta>0$ when $\min\{\sigma_{\boldsymbol{x}},\sigma_{\boldsymbol{y}},\sigma_{\boldsymbol{z}}\}\geq\sqrt{3q_1^2T+4q_2^2\left\lfloor\frac{T}{K}\right\rfloor}\frac{2\sqrt{\beta_0\log(1/\delta)}}{\epsilon}$.

Proof:

According to Theorem 1 in [1], given sampling rate q, noise multiplier σ, and the number of iterations T, the log moment satisfies:

$$\alpha(\lambda)\leq\beta_0\cdot\lambda\cdot T\frac{q^2}{\sigma^2}.\tag{103}$$

In the proposed framework, there are two stages that involve the accumulation of the privacy budget. The first occurs at each iteration, where the workers compute and transmit $\Delta_{\boldsymbol{x}}^{i,t}, \Delta_{\boldsymbol{y}}^{i,t}, \Delta_{\boldsymbol{z}}^{i,t}$ to the master according to Eqs. (13)-(15). In this step, according to the composition of moments (Theorem 2 in [1]), we can get:

$$\begin{aligned} &\alpha_{\mathcal{M}_1}(\lambda) \\ &\leq \alpha_{\mathcal{M}_{\boldsymbol{x}}}(\lambda) + \alpha_{\mathcal{M}_{\boldsymbol{y}}}(\lambda) + \alpha_{\mathcal{M}_{\boldsymbol{z}}}(\lambda) \\ &\leq T \cdot \beta_0 \frac{q_1^2}{\sigma_{\boldsymbol{x}}^2}\lambda^2 + T \cdot \beta_0 \frac{q_1^2}{\sigma_{\boldsymbol{y}}^2}\lambda^2 + T \cdot \beta_0 \frac{q_1^2}{\sigma_{\boldsymbol{z}}^2}\lambda^2. \end{aligned} \tag{104}$$

In the second stage, the value function and cutting plane are updated every K iterations. Specifically, the value function is updated based on the third-level optimization problem, which requires computing and transmitting $\tilde{\Delta}_{\boldsymbol{z}}^{i,t}$ according to Eq. (5), thus, we can obtain:

$$\alpha_{\mathcal{M}_3}(\lambda) \leq \left\lfloor \frac{T}{K} \right\rfloor \cdot \beta_0 \frac{q_2^2}{\sigma_{\boldsymbol{z}}^2}\lambda^2. \tag{105}$$

Then, the outer-level cutting planes are updated to refine the relaxation, which requires computing and transmitting $\tilde{\Delta}_{\boldsymbol{y}}^{i,t}$ according to Eq. (8), it follows that:

$$\alpha_{\mathcal{M}_2}(\lambda) \leq \left\lfloor \frac{T}{K} \right\rfloor \cdot \beta_0 \frac{q_2^2}{\sigma_{\boldsymbol{y}}^2}\lambda^2. \tag{106}$$

Additionally, the Hessian-vector products HVX_i^t and HVZ_i^t will be computed and sent, we have:

$$\alpha_{\mathcal{M}_h}(\lambda) \leq \left\lfloor \frac{T}{K} \right\rfloor \cdot \beta_0 \frac{q_2^2}{\sigma_{\boldsymbol{x}}^2}\lambda^2 + \left\lfloor \frac{T}{K} \right\rfloor \cdot \beta_0 \frac{q_2^2}{\sigma_{\boldsymbol{z}}^2}\lambda^2. \tag{107}$$

Consequently, integrating the composition of moments with Eqs. (104), (105), and (106) yields:

$$\begin{aligned} \alpha_{\mathcal{M}}(\lambda) \leq\ & T \cdot \beta_0 \frac{q_1^2}{\sigma_{\boldsymbol{x}}^2}\lambda^2 + T \cdot \beta_0 \frac{q_1^2}{\sigma_{\boldsymbol{y}}^2}\lambda^2 + T \cdot \beta_0 \frac{q_1^2}{\sigma_{\boldsymbol{z}}^2}\lambda^2 \\ &+ \left\lfloor \frac{T}{K} \right\rfloor \cdot \beta_0 \frac{q_2^2}{\sigma_{\boldsymbol{x}}^2}\lambda^2 + 2\left\lfloor \frac{T}{K} \right\rfloor \cdot \beta_0 \frac{q_2^2}{\sigma_{\boldsymbol{z}}^2}\lambda^2 + \left\lfloor \frac{T}{K} \right\rfloor \cdot \beta_0 \frac{q_2^2}{\sigma_{\boldsymbol{y}}^2}\lambda^2. \end{aligned} \tag{108}$$

According to the tail bound conversion in Theorem 2 of [1], to ensure that the proposed algorithm achieves (ϵ, δ)-differential privacy, it suffices to satisfy:

$$\beta_0\lambda^2 \left(T \cdot \frac{q_1^2}{\sigma_{\boldsymbol{x}}^2} + T \cdot \frac{q_1^2}{\sigma_{\boldsymbol{y}}^2} + T \cdot \frac{q_1^2}{\sigma_{\boldsymbol{z}}^2} + \left\lfloor \frac{T}{K} \right\rfloor \cdot \frac{q_2^2}{\sigma_{\boldsymbol{x}}^2} + 2\left\lfloor \frac{T}{K} \right\rfloor \cdot \frac{q_2^2}{\sigma_{\boldsymbol{z}}^2} + \left\lfloor \frac{T}{K} \right\rfloor \cdot \frac{q_2^2}{\sigma_{\boldsymbol{y}}^2} \right) \leq \lambda\epsilon/2, \tag{109}$$

$$\exp(-\lambda\epsilon/2) \leq \delta. \tag{110}$$

In addition, according to Theorem 1 in [1], we need:

$$\lambda \leq \min\{\sigma_{\boldsymbol{x}}^2 \log(\frac{1}{\max\{q_1, q_2\}\sigma_{\boldsymbol{x}}}), \sigma_{\boldsymbol{y}}^2 \log(\frac{1}{\max\{q_1, q_2\}\sigma_{\boldsymbol{y}}}), \sigma_{\boldsymbol{z}}^2 \log(\frac{1}{\max\{q_1, q_2\}\sigma_{\boldsymbol{z}}})\}. \tag{111}$$

Given a constant $\beta_1 \leq \log(\frac{1}{\max\{q_1,q_2\}\cdot\max\{\sigma_{\boldsymbol{x}},\sigma_{\boldsymbol{y}},\sigma_{\boldsymbol{z}}\}})$, we have that if λ satisfies:

$$\lambda \leq \beta_1 \min\{\sigma_{\boldsymbol{x}}^2, \sigma_{\boldsymbol{y}}^2, \sigma_{\boldsymbol{z}}^2\}, \tag{112}$$

then it also satisfies Eq. (111). Defining $\hat{\sigma}^2 = \min\{\sigma_{\boldsymbol{x}}^2, \sigma_{\boldsymbol{y}}^2, \sigma_{\boldsymbol{z}}^2\}$, we can also derive the following inequality:

$$T\left(\frac{q_1^2}{\sigma_{\boldsymbol{x}}^2} + \frac{q_1^2}{\sigma_{\boldsymbol{y}}^2} + \frac{q_1^2}{\sigma_{\boldsymbol{z}}^2}\right) + \left\lfloor \frac{T}{K} \right\rfloor \cdot \left(\frac{q_2^2}{\sigma_{\boldsymbol{x}}^2} + \frac{2q_2^2}{\sigma_{\boldsymbol{z}}^2} + \frac{q_2^2}{\sigma_{\boldsymbol{y}}^2}\right) \leq \frac{3q_1^2T}{\hat{\sigma}^2} + \frac{4q_2^2\left\lfloor \frac{T}{K} \right\rfloor}{\hat{\sigma}^2}. \tag{113}$$

By setting $\lambda = \frac{2\log(1/\delta)}{\epsilon}$, we can get:

$$\frac{2\beta_0\log(1/\delta)}{\epsilon}\left(T \cdot \frac{q_1^2}{\sigma_{\boldsymbol{x}}^2} + T \cdot \frac{q_1^2}{\sigma_{\boldsymbol{y}}^2} + T \cdot \frac{q_1^2}{\sigma_{\boldsymbol{z}}^2} + \left\lfloor \frac{T}{K} \right\rfloor \cdot \frac{q_2^2}{\sigma_{\boldsymbol{x}}^2} + 2\left\lfloor \frac{T}{K} \right\rfloor \cdot \frac{q_2^2}{\sigma_{\boldsymbol{z}}^2} + \left\lfloor \frac{T}{K} \right\rfloor \cdot \frac{q_2^2}{\sigma_{\boldsymbol{y}}^2} \right) \leq \frac{\epsilon}{2}, \tag{114}$$

that is,

$$T\left(\frac{q_1^2}{\sigma_{\boldsymbol{x}}^2}+\frac{q_1^2}{\sigma_{\boldsymbol{y}}^2}+\frac{q_1^2}{\sigma_{\boldsymbol{z}}^2}\right)+\left\lfloor\frac{T}{K}\right\rfloor\cdot\left(\frac{q_2^2}{\sigma_{\boldsymbol{x}}^2}+\frac{2q_2^2}{\sigma_{\boldsymbol{z}}^2}+\frac{q_2^2}{\sigma_{\boldsymbol{y}}^2}\right)\leq\frac{\epsilon^2}{4\beta_0\log(1/\delta)}. \tag{115}$$

To guarantee that the Eq. (115) holds for all cases, let $\frac{3q_1^2T}{\hat{\sigma}^2}+\frac{4q_2^2\lfloor\frac{T}{K}\rfloor}{\hat{\sigma}^2}\leq\frac{\epsilon^2}{4\beta_0\log(1/\delta)}$, which yields:

$$\hat{\sigma}^2\geq\left(3q_1^2T+4q_2^2\left\lfloor\frac{T}{K}\right\rfloor\right)\frac{4\beta_0\log(1/\delta)}{\epsilon^2}. \tag{116}$$

By combining the setting $\lambda=\frac{2\log(1/\delta)}{\epsilon}$ with Eq. (116), we can derive that for any privacy budget

$$\epsilon\leq 2\beta_0\beta_1\left(3q_1^2T+4q_2^2\left\lfloor\frac{T}{K}\right\rfloor\right), \tag{117}$$

the Eq. (111) is always satisfied. Consequently, we can conclude that there exist constants β_0,β_1 for any privacy budget $\epsilon\leq 2\beta_0\beta_1\left(3q_1^2T+4q_2^2\lfloor\frac{T}{K}\rfloor\right)$ with sampling rates $q_1=m/M$ and $q_2=\tilde{m}/M$, when $\min\{\sigma_{\boldsymbol{x}},\sigma_{\boldsymbol{y}},\sigma_{\boldsymbol{z}}\}\geq\sqrt{3q_1^2T+4q_2^2\lfloor\frac{T}{K}\rfloor}\frac{2\sqrt{\beta_0\log(1/\delta)}}{\epsilon}$, the proposed algorithm is (ϵ,δ)-differentially private for any $\delta>0$.

□

Corollary F.1 ***(Tripartite Trade-off among Privacy, Utility, and Complexity)*** *To guarantee that the proposed DPDTO satisfies (ϵ,δ)-differential privacy as defined in Theorem 4.5, we set the noise multipliers as $\sigma_{\boldsymbol{x}}^2=\sigma_{\boldsymbol{y}}^2=\sigma_{\boldsymbol{z}}^2=\frac{4\beta_0\log(1/\delta)}{\epsilon^2}\left(3q_1^2T+4q_2^2\lfloor\frac{T}{K}\rfloor\right)$. By substituting these values into the convergence bound in Theorem 4.4, we obtain the non-asymptotic convergence rate with the privacy guarantee as follows:*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla L_\lambda(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t)\|^2\leq\frac{4(L_\lambda(\boldsymbol{x}_0,\boldsymbol{y}_0,\boldsymbol{z}_0)-f^*)}{\sqrt{T}}+\left(\frac{2G_{\boldsymbol{x}}^4}{c_{\boldsymbol{x}}^2}+\frac{4G_{\boldsymbol{y}}^4}{c_{\boldsymbol{y}}^2}+\frac{4G_{\boldsymbol{z}}^4}{c_{\boldsymbol{z}}^2}\right)\\
&+\frac{2\hat{L}^2(c_{\boldsymbol{x}}^2+c_{\boldsymbol{y}}^2+c_{\boldsymbol{z}}^2)}{\sqrt{T}}+\frac{\hat{L}^2(18c_{\boldsymbol{x}}^2+12c_{\boldsymbol{y}}^2)}{T}\\
&+\frac{\hat{L}^2\beta_0\log(1/\delta)\left(3q_1^2T+4q_2^2\lfloor\frac{T}{K}\rfloor\right)}{\epsilon^2\sqrt{T}nm^2}\left(32d_1c_{\boldsymbol{x}}^2+32d_2c_{\boldsymbol{y}}^2+32d_3c_{\boldsymbol{z}}^2\right)\\
&+\frac{\hat{L}^2\beta_0\log(1/\delta)\left(3q_1^2T+4q_2^2\lfloor\frac{T}{K}\rfloor\right)}{\epsilon^2T}\left(\frac{96d_1c_{\boldsymbol{x}}^2+64d_2c_{\boldsymbol{y}}^2}{nm^2}+\frac{192d_1c_{\boldsymbol{x}}^2+128d_2c_{\boldsymbol{y}}^2}{m^2}\right).
\end{aligned} \tag{118}$$

Discussion: Corollary F.1 reveals a fundamental *tripartite trade-off* in the proposed DPDTO among privacy, optimization utility, and communication complexity. First, as indicated by the last two terms of Eq. (118), enforcing stronger privacy guarantees (corresponding to a smaller ϵ) requires injecting more noise, thereby slowing the convergence rate and impairing final performance. Second, the parameter K governs the balance between communication complexity and optimization utility. While a smaller K provides a tighter value-function-embedded cutting plane relaxation and better performance, a larger K reduces communication overhead (as demonstrated in Theorem 4.6). However, a larger K also introduces stale value functions and outdated cutting planes, negatively impacting the overall optimization utility. This complex dynamic underscores a *tripartite trade-off* among privacy guarantees, optimization utility, and communication complexity.

G Proof of Theorem 4.6

Theorem 4.6 (Communication Complexity of DPDTO) The communication complexity is typically measured as the total number of bits exchanged throughout the entire iterative process [84, 50]. Accordingly, the communication complexity of the proposed method can be formulated as:

$$64n(d_1+d_2+d_3)T+32n(d_1+2d_2+3d_3)\lfloor T/K\rfloor. \tag{119}$$

Proof:

In the proposed method, there are two primary processes involving information exchange: the construction and update of the value-function-embedded privacy-preserving cutting plane relaxation, and the iterative update of variables within the privacy-preserving distributed optimization algorithm. In the first process, at every K iteration, worker i first transmits $\tilde{\Delta}_{\boldsymbol{z}}^{i,t} \in \mathbb{R}^{d_3}$ to the master. Then, the master computes and broadcasts the updated $\boldsymbol{z}^* = \boldsymbol{z}^t - \frac{\eta_{\boldsymbol{z}}}{n} \sum_{i=1}^n \tilde{\Delta}_{\boldsymbol{z}}^t \in \mathbb{R}^{d_3}$ back to the workers. Therefore, the communication complexity of this stage can be expressed as:

$$\mathcal{I}_1 = 64nd_3 \left\lfloor \frac{T}{K} \right\rfloor. \tag{120}$$

Next, the algorithm proceeds to update the outer-level cutting planes. During this phase, each worker first transmits $\tilde{\Delta}_{\boldsymbol{y}}^{i,t} \in \mathbb{R}^{d_2}$ to the master node. The master then computes $\bar{\Delta}^t = \frac{2\eta_{\boldsymbol{y}}}{n} \hat{V}(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t) \in \mathbb{R}^{d_2}$ and broadcasts it back to all workers. Following this, the workers calculate their respective local Hessian-vector products, $HVX_i^t \in \mathbb{R}^{d_1}$ and $HVZ_i^t \in \mathbb{R}^{d_3}$, and upload them to the master node to construct the new cutting planes. Consequently, the communication complexity for this stage is given by

$$\mathcal{I}_2 = (32nd_1 + 64nd_2 + 32nd_3) \left\lfloor \frac{T}{K} \right\rfloor. \tag{121}$$

Regarding the second procedure, at each iteration, each worker i uploads its local updates, that is, $\Delta_{\boldsymbol{x}}^{i,t} \in \mathbb{R}^{d_1}$, $\Delta_{\boldsymbol{y}}^{i,t} \in \mathbb{R}^{d_2}$, and $\Delta_{\boldsymbol{z}}^{i,t} \in \mathbb{R}^{d_3}$, to the master node. The master then aggregates this received information to update the global variables $\boldsymbol{x}^{t+1} \in \mathbb{R}^{d_1}$, $\boldsymbol{y}^{t+1} \in \mathbb{R}^{d_2}$, and $\boldsymbol{z}^{t+1} \in \mathbb{R}^{d_3}$, which are subsequently transmitted back to all workers. Thus, the communication complexity for this phase is

$$\mathcal{I}_3 = 64n(d_1 + d_2 + d_3)T. \tag{122}$$

Consequently, according to Eqs. (120), (121), and (122), we can conclude that the overall communication complexity of the proposed method can be expressed as:

$$\mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3 = 64n(d_1 + d_2 + d_3)T + 32n(d_1 + 2d_2 + 3d_3) \left\lfloor \frac{T}{K} \right\rfloor. \tag{123}$$

□

H Limitations

In this work, we propose a distributed trilevel optimization framework with theoretical privacy guarantees. To the best of our knowledge, this is the first work to investigate differentially private trilevel optimization. As discussed in this work, trilevel optimization is more vulnerable to privacy leakage than single-level and bilevel optimization. Similarly, multilevel optimization problems with more than three levels are even more susceptible to privacy leakage. However, since such problems are relatively less common in machine learning applications, they are not considered in this work and are left for future investigation.

I Broader Impacts

This work has positive societal impacts by enabling efficient and privacy-preserving solutions to distributed trilevel optimization problems, which are widely used in machine learning and networking. The proposed method facilitates the secure, scalable, and collaborative deployment of AI systems in highly sensitive domains, such as healthcare and finance, thereby effectively mitigating the risks of data leakage.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction are supported by the proposed algorithm in Sec. 3, the theoretical analyses in Sec. 4, and the experimental results in Sec. 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of this work are discussed in Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All assumptions are shown in Sec. 4. The complete proofs of the theoretical results are provided in Appendix D, E, F, and G.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The detailed experimental settings are provided in Sec. 5 and Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: The principal contribution of this study lies in the theoretical realm.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to Appendix A.1 and A.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All experiments in this paper were repeated five times. Please refer to the experimental results in Sec. 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Detailed information about the computational resources and experimental settings is provided in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes. The research complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The societal impacts of this work are discussed in Appendix I.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, all datasets are properly cited and the licenses are respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The proposed method does not involve LLMs as a component of the core methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.