OPTIMAL MOMENTUM METHODS FOR STOCHASTIC MULTILEVEL COMPOSITIONAL OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper investigates stochastic multi-level optimization where the objective is a nested composition of several smooth non-convex functions. We assume that only stochastic estimates of the gradient and function values for each level are accessible. Consequently, obtaining an accurate estimate of the overall gradient is challenging due to the nested structure. To address this, we employ a momentum-based estimator with mini-batches to track the function values of each level, which are subsequently used to construct momentum gradient estimators. We establish an optimal sample complexity of $\mathcal{O}(\epsilon^{-4})$ for finding an ϵ-stationary point, avoiding the stronger average smoothness assumption commonly relied upon in prior literature. Furthermore, by employing a normalization technique, we attain the same rate without requiring problem-dependent constants to set hyperparameters. To achieve the optimal rate without mini-batches, we further develop a batch-free method that incorporates a first-order approximation and a clipping technique for function value estimation. Finally, we validate the effectiveness of our proposed methods through experiments on risk-averse portfolio optimization and hierarchical tilted empirical risk minimization.

1 INTRODUCTION

In this paper, we study stochastic multi-level compositional optimization (Yang et al., 2019; Chen et al., 2021; Zhang & Xiao, 2021; Balasubramanian et al., 2022; Jiang et al., 2025) of the form

$$\min_{\mathbf{x}} F(\mathbf{x}) = f_T \circ \cdots \circ f_1(\mathbf{x}). \tag{1}$$

Here, the objective is a nested composition of T smooth and non-convex functions. We assume that only unbiased stochastic estimates are available for the function values and gradients at each level i, denoted as $f_i(\cdot;\xi_i)$ and $\nabla f_i(\cdot;\xi_i)$, such that $\mathbb{E}_{\xi_i}\left[f_i(\cdot;\xi_i)\right] = f_i(\cdot)$ and $\mathbb{E}_{\xi_i}\left[\nabla f_i(\cdot;\xi_i)\right] = \nabla f_i(\cdot)$.

Multi-level optimization enjoys broad applications in machine learning, including risk-averse optimization (Bruno et al., 2016; Shapiro et al., 2021), reinforcement learning (Dann et al., 2014), model-agnostic meta-learning (Ji et al., 2020), graph neural network training (Cong et al., 2020), stochastic variational inequalities (Ruszczyński, 2021), adversarial attacks and defenses in deep neural networks (Yuan et al., 2019), risk management (Cole et al., 2017; Dentcheva et al., 2017), and hierarchical tilted empirical risk minimization (Li et al., 2021; 2023). Unlike traditional single-level stochastic optimization, where unbiased stochastic gradient evaluations are readily available, obtaining an unbiased estimate of the overall gradient and function value is generally infeasible in multi-level problems. This difficulty stems from the nested composition structure, specifically:

$$\mathbb{E}_{\xi_T,\cdots,\xi_2,\xi_1}[f_T(\cdots f_2(f_1(\mathbf{x};\xi_1);\xi_2)\cdots;\xi_T)] \neq \mathbb{E}_{\xi_T}[f_T(\cdots \mathbb{E}_{\xi_2}[f_2(\mathbb{E}_{\xi_1}[f_1(\mathbf{x};\xi_1)];\xi_2)]\cdots;\xi_T)].$$

Furthermore, estimation errors can accumulate as the composition depth increases, making the problem more challenging. To improve sample complexity, existing literature often imposes additional *average smoothness* and *average Lipschitz* conditions (Chen et al., 2021; Zhang & Xiao, 2021; Jiang et al., 2025), which require that stochastic estimates of each layer be smooth and Lipschitz continuous. However, in single-level stochastic optimization, it is well known that momentum-based methods can achieve the optimal sample complexity of $\mathcal{O}(\epsilon^{-4})$ under standard smoothness assumptions (rather than the stronger average smoothness condition). This motivates the following question: *Is it possible to obtain the optimal rate in stochastic multi-level optimization without these additional assumptions?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this work, we answer this question affirmatively by proving the optimal $\mathcal{O}(\epsilon^{-4})$ sample complexity using a simple momentum-based approach with mini-batches. Our analysis avoids the restrictive average smoothness and Lipschitz continuous conditions. Furthermore, we provide a normalized variant that eliminates the need for prior knowledge of problem-dependent constants when setting hyperparameters. Finally, to avoid large batches, we develop a batch-free method that leverages first-order approximation together with clipping, preserving the optimal sample complexity with a constant batch size. To summarize, our main contributions are as follows:

- Our proposed MML method uses momentum-based estimators to track function values and gradients, obtaining the optimal $\mathcal{O}(\epsilon^{-4})$ sample complexity with mini-batches. Our analysis does not rely on the average smoothness and Lipschitz continuity used in previous literature.
- Employing a normalization technique, our MML-v2 removes the reliance on problem-dependent constants to set hyperparameters while preserving optimal sample complexity.
- To avoid large batches, the proposed MMLFA method uses a first-order approximation for function-value tracking, combined with a clipping operation to control the estimator range. This design allows us to maintain the optimal rate with a constant batch size.
- We validate the effectiveness of the proposed methods through numerical experiments on risk-averse portfolio optimization and hierarchical tilted empirical risk minimization.

2 Related Work

In this section, we review related work on stochastic multi-level optimization and its decentralized/federated variants.

2.1 Multi-level Optimization

Multi-level compositional optimization was first investigated by Yang et al. (2019), who developed stochastic methods based on an extrapolation-interpolation scheme. They established a sample complexity of $\mathcal{O}(\epsilon^{-(7+T)/2})$ for non-convex objectives and $\mathcal{O}(\epsilon^{-(3+T)/4})$ for strongly convex objectives. To further improve the convergence rate, Chen et al. (2021) proposed the stochastically corrected stochastic compositional gradient method, deriving a sample complexity of $\mathcal{O}(\epsilon^{-4})$. However, their analysis requires the stronger average smoothness assumption, which is typically used to obtain improved $\mathcal{O}(\epsilon^{-3})$ rates via variance reduction. Concurrently, Balasubramanian et al. (2022) obtained $\mathcal{O}(\epsilon^{-6})$ sample complexity for a momentum-based method with mini-batches and improved the rate to $\mathcal{O}(\epsilon^{-4})$ using linearized stochastic estimates. However, their analysis neglects dependency issues, rendering the results invalid (a detailed discussion is provided in the appendix). To enhance the rate under the average smoothness assumption, Zhang & Xiao (2021) introduced the normalized proximal approximate gradient method, which evaluates gradients via nested variance reduction with large batch sizes. They achieved a total sample complexity of $\mathcal{O}(\epsilon^{-3})$ in the expectation setting, and $\mathcal{O}(n + \sqrt{n}\epsilon^{-2})$ for the finite-sum structure, where n denotes the number of component functions across all levels. Similar results were obtained by Jiang et al. (2022; 2025) using a constant batch size, further deriving optimal rates for convex and strongly convex functions.

Beyond smooth non-convex settings, Ruszczyński (2021) considered non-smooth non-convex multi-level optimization, assuming that functions at each level are locally Lipschitz continuous and differentiable in a generalized sense. For smooth and convex problems, Zhang & Lan (2025) proposed a stochastic sequential dual method, achieving a sample complexity of $\mathcal{O}(\epsilon^{-2})$. Constrained multi-level optimization has also been studied recently. Xiao et al. (2022) proposed a projection-free conditional gradient-type algorithm, attaining $\mathcal{O}(\epsilon^{-2})$ and $\mathcal{O}(\epsilon^{-3})$ complexities for stochastic first-order and linear-minimization oracles. Following this, Jiang et al. (2024) presented a comprehensive study on projection-free methods, covering non-convex, convex, and strongly convex objectives. Recent studies have also focused on analyzing stability and generalization performance (Pan et al., 2024) and extending multi-level optimization to minimax settings (Zhang & Gao, 2025).

2.2 Decentralized/Federated Multi-level Optimization

Multi-level optimization has also been widely studied in decentralized and federated environments. Gao (2024b) first studied decentralized stochastic multi-level optimization, proposing two algorithms

that achieve level-independent convergence rates for non-convex problems under milder conditions. For smooth and strongly convex objectives, Yang & Li (2024) proposed the federated stochastic compositional gradient extrapolation method. Based on a double-loop strategy, this method attains an optimal $\mathcal{O}(\log(1/\epsilon))$ communication complexity while maintaining an almost optimal $\widetilde{\mathcal{O}}(1/\epsilon)$ sample complexity. To address imbalanced data classification in federated learning, Zhang et al. (2024) developed a federated multi-level compositional minimax algorithm for AUC maximization, equipped with rigorous theoretical guarantees. To achieve linear speedup under heterogeneous settings, Gao (2024a) developed a federated stochastic multi-level algorithm based on a Jacobian-vector product estimator and proved level-independent convergence for non-convex objectives.

3 Momentum-based Multi-level Algorithm

In this section, we first outline the assumptions used and then present the proposed momentum-based methods, followed by their convergence guarantees.

3.1 Assumptions

We adopt the following assumptions, which are standard in stochastic multi-level optimization (Yang et al., 2019; Chen et al., 2021; Zhang & Xiao, 2021; Balasubramanian et al., 2022; Jiang et al., 2025).

Assumption 3.1. Denote $F_\star = \min_{\mathbf{x}} F(\mathbf{x})$ and suppose $F\left(\mathbf{x}^1\right) - F_\star \leq \Delta_F$ for the initial point $\mathbf{x}^1$.

Assumption 3.2. All functions $f_1, \ldots, f_T$ are L_f-Lipschitz continuous, and their gradients $\nabla f_1, \ldots, \nabla f_T$ are also L_J-Lipschitz continuous.

Remark: Under Assumption 3.2, the objective F is L_F-smooth, where $L_F \coloneqq L_f^{2T-1} L_J \sum_{i=1}^{T} L_f^{-i}$.

Assumption 3.3. The function and gradient estimates are unbiased with bounded variance, i.e.,

$$\mathbb{E}_\xi \left[f_i(\mathbf{x};\xi)\right] = f_i(\mathbf{x}), \quad \mathbb{E}_\xi \left[\nabla f_i(\mathbf{x};\xi)\right] = \nabla f_i(\mathbf{x}),$$
$$\mathbb{E}_\xi \left[\left\|f_i(\mathbf{x};\xi) - f_i(\mathbf{x})\right\|^2\right] \leq \sigma^2, \quad \mathbb{E}_\xi \left[\left\|\nabla f_i(\mathbf{x};\xi) - \nabla f_i(\mathbf{x})\right\|^2\right] \leq \sigma^2.$$

Remark: These assumptions are standard in traditional stochastic compositional optimization. Some existing multi-level literature (Chen et al., 2021; Zhang & Xiao, 2021; Jiang et al., 2025) requires the stronger average smoothness and Lipschitz condition, namely, for each $1 \leq i \leq T$:

$$\mathbb{E}_\xi \left[\left\|f_i(\mathbf{x};\xi) - f_i(\mathbf{y};\xi)\right\|^2\right] \leq L_f^2 \left\|\mathbf{x} - \mathbf{y}\right\|^2, \mathbb{E}_\xi \left[\left\|\nabla f_i(\mathbf{x};\xi) - \nabla f_i(\mathbf{y};\xi)\right\|^2\right] \leq L_J^2 \left\|\mathbf{x} - \mathbf{y}\right\|^2.$$

However, this assumption is not required in our work.

3.2 The Proposed Method

We now present our algorithm for stochastic multi-level optimization. In traditional stochastic non-convex optimization (Kingma & Ba, 2015; Cutkosky & Orabona, 2019), we can directly access an unbiased stochastic gradient estimate of the overall objective $\nabla F(\cdot)$. However, this is no longer possible in multi-level optimization, even though unbiased estimates of the function and gradient at each individual layer are available. Since the overall gradient $\nabla F(\mathbf{x})$ typically involves nested expectations, multiplying the stochastic gradients of each layer via the chain rule, i.e., $\prod_{i=1}^{T} \nabla f_i(\cdot;\xi_i)$, yields only a biased estimation, potentially resulting in significant estimation error. This limitation arises from the nested nature of the problem, since $\mathbb{E}_{\xi,\zeta}[f_2(f_1(\mathbf{x};\zeta);\xi)] \neq \mathbb{E}_\xi[f_2(\mathbb{E}_\zeta[f_1(\mathbf{x};\zeta)]);\xi]$.

To address this issue, we first provide a detailed analysis of the gradient estimation error. For clarity, we consider the simple two-level case where $T = 2$. The objective function is $F(\mathbf{x}) = \mathbb{E}_\xi[f_2(\mathbb{E}_\zeta[f_1(\mathbf{x};\zeta)];\xi)]$, and its gradient is $\nabla F(\mathbf{x}) = \mathbb{E}_\xi[\nabla f_2(\mathbb{E}_\zeta[f_1(\mathbf{x};\zeta)];\xi)] \cdot \mathbb{E}_\omega[\nabla f_1(\mathbf{x};\omega)]$. Note that there exist three distinct expectations $\mathbb{E}_\xi, \mathbb{E}_\zeta, \mathbb{E}_\omega$. To construct an effective gradient estimator, we must derive accurate estimates of the gradients $\mathbb{E}_\xi[\nabla f_2(\cdot;\xi)] \cdot \mathbb{E}_\omega[\nabla f_1(\cdot;\omega)]$ and the inner function values $\mathbb{E}_\zeta[f_1(\cdot;\zeta)]$. Therefore, we first employ a momentum-based estimator to track the function values of each layer i at time step k:

$$\mathbf{u}_i^k = (1-\alpha)\mathbf{u}_i^{k-1} + \alpha \frac{1}{B_1} \sum_{j=1}^{B_1} f_i(\mathbf{u}_{i-1}^k; \xi_{i,j}^k), \tag{2}$$

Algorithm 1 MML method

1: **Input:** parameters η, α
2: **for** time step $k = 1$ **to** K **do**
3: Set $\mathbf{u}_0^k = \mathbf{x}^k$
4: Compute the function estimator according to equation (2)
5: Compute the gradient estimator $\mathbf{z}^k$ according to equation (3)
6: (v1) Update the weight: $\mathbf{x}^{k+1} = \mathbf{x}^k - \eta \mathbf{z}^k$
7: (v2) Update the weight: $\mathbf{x}^{k+1} = \mathbf{x}^k - \eta \frac{\mathbf{z}^k}{\|\mathbf{z}^k\|}$
8: **end for**

where $\mathbf{u}_0^k = \mathbf{x}^k$ and $\mathbf{u}_i^1 = \frac{1}{B_0}\sum_{j=1}^{B_0} f_i(\mathbf{u}_{i-1}^1; \xi_{i,j}^1)$. Note that multi-level problems are more difficult than two-level problems, because the estimation errors may accumulate exponentially with the composition depth. This is why we employ the level-wise function estimators $\mathbf{u}_i^k$. We analyze the coupled recursions for Γ_i and Λ_i in the following lemma, where we denote the error of the function estimator as $\Gamma_i = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\left\|\mathbf{u}_i^k - f_i(\mathbf{u}_{i-1}^k)\right\|^2\right]$ and the difference between adjacent steps as $\Lambda_i = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\left\|\mathbf{u}_i^{k+1} - \mathbf{u}_i^k\right\|^2\right]$. We will then use them to control the gradient estimator error.

Lemma 3.4. *Set $B_1 \leq \alpha^2 B_0 K$. We can ensure that*

$$\Gamma_i \leq \frac{2L_f^2}{\alpha^2}\Lambda_{i-1} + \frac{2\alpha\sigma^2}{B_1}; \quad \Lambda_i \leq 2\alpha^2 L_f^2 \Lambda_{i-1} + \frac{\alpha^2\sigma^2}{B_1} + 2\alpha^2\Gamma_i.$$

The function estimation error Γ_i and the step difference Λ_i are mutually dependent. We can bound them simultaneously in the following lemma.

Lemma 3.5. *Let $C_{i+1} = 5 + 6L_f^2 C_i$ with $C_0 = 1$. Then*

$$\Gamma_i \leq C_i\left(\frac{1}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + \frac{\sigma^2}{B_1}\right);$$
$$\Lambda_i \leq C_i\left(\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + \frac{\alpha^2\sigma^2}{B_1}\right).$$

Remark: By properly setting the learning rate, we can control the term $\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right]$. Consequently, using a sufficiently large batch size B_1 yields a small function-value estimation error.

Using the function value estimators $\mathbf{u}_i^k$, we develop the overall gradient estimator via the chain rule:

$$\mathbf{z}^k = (1-\alpha)\mathbf{z}^{k-1} + \alpha\frac{1}{B_1}\sum_{j=1}^{B_1}\prod_{i=1}^{T}\nabla f_i(\mathbf{u}_{i-1}^k; \xi_{i,j}^k), \tag{3}$$

where $\mathbf{z}^1 = \frac{1}{B_0}\sum_{j=1}^{B_0}\prod_{i=1}^{T}\nabla f_i(\mathbf{u}_{i-1}^1; \xi_{i,j}^1)$. This gradient estimator also follows a momentum update, and its error is bounded as follows.

Lemma 3.6. *We have the following guarantee:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{z}^k - \nabla F(\mathbf{x}^k)\right\|^2\right] \leq \frac{2\alpha}{B_1}T^2(\sigma^2 + L_f^2)^T + \frac{\sigma^2}{B_1}4TL_F^2\sum_{i=1}^{T-1}C_i$$
$$+ \left(1 + T\sum_{i=1}^{T-1}C_i\right)\frac{4L_F^2}{\alpha^2 K}\sum_{k=1}^{K+1}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right].$$

Remark: With suitable choices of the learning rate η and momentum parameter α, we can ensure a small gradient estimation error using a large batch size B_1.

Once the gradient estimator is obtained, the decision variable is updated using either standard SGD (i.e., $\mathbf{x}^{k+1} = \mathbf{x}^k - \eta\mathbf{z}^k$) or normalized SGD (i.e., $\mathbf{x}^{k+1} = \mathbf{x}^k - \eta\frac{\mathbf{z}^k}{\|\mathbf{z}^k\|}$). The complete procedure,

named the Momentum-based Multi-Level method (MML), is presented in Algorithm 1. We now present the convergence guarantees below.

Theorem 3.7. *Under Assumptions 3.1, 3.2, 3.3, by setting $B_0 = \mathcal{O}(1)$, $B_1 = \mathcal{O}(K)$, $\eta = \mathcal{O}(1)$ and $\alpha = \mathcal{O}(1)$, our MML-v1 algorithm ensures*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}^k)\right\|\right] \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right).$$

Remark: To ensure that $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}^k)\right\|\right] \leq \epsilon$, the iteration complexity is $K = \mathcal{O}(\epsilon^{-2})$ and the overall sample complexity is $B_1 K = \mathcal{O}(\epsilon^{-4})$. This matches the lower bound for stochastic single-level optimization under standard smoothness assumptions (Arjevani et al., 2019) and is therefore optimal. Notably, this is the first result to obtain optimal $\mathcal{O}(\epsilon^{-4})$ sample complexity for multi-level optimization under standard smoothness assumptions.

While Theorem 3.7 attains optimal sample complexity, it requires knowledge of problem-dependent constants to set hyperparameters, e.g., $\eta \leq \min\{\frac{1}{2L_F}, \frac{\alpha}{(4L_F\sqrt{1+T\sum_{i=1}^{T-1}C_i})}\}$. We next show that this requirement can be removed using the normalization technique, as shown in the following theorem.

Theorem 3.8. *Under Assumptions 3.1, 3.2, 3.3, by setting $B_0 = 1$, $B_1 = \lfloor K/4 \rfloor$, $\eta = K^{-1/2}$, and $\alpha = 1/2$, our MML-v2 algorithm ensures*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}^k)\right\|\right] \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right).$$

Remark: Theorem 3.8 provides the same rate while avoiding the use of problem-dependent constants for setting hyperparameters. Note that $\alpha = 1/2$ is not essential to the analysis and can be any positive constant in the range $(0, 1)$. A similar normalization technique is also used by Hübler et al. (2024) in single-level problems. However, our analysis is different from Hübler et al. (2024), which analyzes with decreasing η_t and increasing α_t, and obtains a nearly optimal rate of $\mathcal{O}(K^{-1/4}\log K)$.

4 Momentum-based Multi-level Optimization with First-order Approximation

In the previous section, we achieved optimal sample complexity using mini-batches. While large batches can be beneficial (e.g., enabling parallelism), they would become impractical when hardware or memory constraints prevent the use of $B_1 = \mathcal{O}(K)$. In this section, we introduce an alternative algorithm that achieves the optimal sample complexity without relying on large batches.

A careful examination of the analysis reveals that the large batch requirement stems from Lemma 3.5, specifically the term σ^2/B_1 in the estimation error. To ensure a small function estimation error, a large B_1 is required, rather than a constant batch size of $B_1 = \mathcal{O}(1)$. To address this issue, we propose to track function values using the first-order Taylor approximation:

$$f_i(\mathbf{u}_{i-1}^k) \approx f_i(\mathbf{u}_{i-1}^{k-1}) + \langle \nabla f_i(\mathbf{u}_{i-1}^k), \mathbf{u}_{i-1}^k - \mathbf{u}_{i-1}^{k-1}\rangle.$$

Recall that the estimator $\mathbf{u}_i^k$ tracks $f_i(\mathbf{u}_{i-1}^k)$ and $\mathbf{u}_i^{k-1}$ estimates $f_i(\mathbf{u}_{i-1}^{k-1})$. Moreover, we can obtain a stochastic estimate of $\nabla f_i(\mathbf{u}_{i-1}^k)$, denoted by $\nabla f_i(\mathbf{u}_{i-1}^k; \xi_i^k)$. Consequently, we approximate

$$\mathbf{u}_i^k \approx \mathbf{u}_i^{k-1} + \langle \nabla f_i(\mathbf{u}_{i-1}^k; \xi_i^k), \mathbf{u}_{i-1}^k - \mathbf{u}_{i-1}^{k-1}\rangle.$$

Since $f_i(\mathbf{u}_{i-1}^k; \xi_i^k)$ is an unbiased estimator of $f_i(\mathbf{u}_{i-1}^k)$, we combine these two estimators to form

$$\tilde{\mathbf{u}}_i^k = \alpha f_i(\mathbf{u}_{i-1}^k; \xi_i^k) + (1-\alpha)\left(\mathbf{u}_i^{k-1} + \langle \nabla f_i(\mathbf{u}_{i-1}^k; \xi_i^k), \mathbf{u}_{i-1}^k - \mathbf{u}_{i-1}^{k-1}\rangle\right), \tag{4}$$

where $\tilde{\mathbf{u}}_i^1 = f_i(\mathbf{u}_{i-1}^1; \xi_i^1)$. The parameter α serves as a trade-off factor between two estimation methods. Although this new estimator requires computing $\nabla f_i(\mathbf{u}_{i-1}^k; \xi_i^k)$, this term is also needed for constructing the gradient estimator. Thus, we can improve the estimator without increasing the overall sample complexity. Compared to the traditional momentum-based estimator, we incorporate the

Algorithm 2 MMLFA method

1: **Input:** parameters η, α, C
2: **for** time step $k = 1$ **to** K **do**
3: Set $\mathbf{u}_0^k = \mathbf{x}^k$
4: Compute the function estimator $\tilde{\mathbf{u}}_i^k$ according to equation (4)
5: Conduct the clipping operation $\mathbf{u}_i^k = \mathrm{Clip}_C[\tilde{\mathbf{u}}_i^k]$
6: Compute the gradient estimator $\mathbf{z}^k$ according to equation (5)
7: Update the weight: $\mathbf{x}^{k+1} = \mathbf{x}^k - \eta \mathbf{z}^k$
8: **end for**

additional term $(1-\alpha)\langle \nabla f_i(\mathbf{u}_{i-1}^k; \xi_i^k), \mathbf{u}_{i-1}^k - \mathbf{u}_{i-1}^{k-1} \rangle$, which compensates for the drift $\mathbf{u}_i^k - \mathbf{u}_i^{k-1}$ via the first-order Taylor expansion. After computing $\tilde{\mathbf{u}}_i^k$, we apply a clipping operation to keep the estimator within a controlled range:

$$\mathbf{u}_i^k = \mathrm{Clip}_C[\tilde{\mathbf{u}}_i^k].$$

This operation is straightforward to implement:

$$\mathbf{u}_i^k = \begin{cases} \tilde{\mathbf{u}}_i^k, & \text{if } \left\|\tilde{\mathbf{u}}_i^k\right\| \le C, \\ \dfrac{C}{\left\|\tilde{\mathbf{u}}_i^k\right\|} \tilde{\mathbf{u}}_i^k, & \text{otherwise.} \end{cases}$$

Here, parameter C is the clipping threshold and we require $\left\|f_i(\mathbf{u}_{i-1}^k)\right\| \le C$ for each i and k. In practice, we simply set C as a large value, as discussed in our experiments. Next, we derive the guarantee for this new estimator. As before, we define that $\Gamma_i = \frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\left\|\mathbf{u}_i^k - f_i(\mathbf{u}_{i-1}^k)\right\|^2\right]$ and $\Lambda_i = \frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\left\|\mathbf{u}_i^{k+1} - \mathbf{u}_i^k\right\|^2\right]$.

Lemma 4.1. *Set $1 \le \alpha^2 K$, and we can ensure that*

$$\Gamma_i \le \frac{(4L_f^2 + 2L_J C + 2\sigma^2)}{\alpha}\Lambda_{i-1} + 3\alpha\sigma^2; \quad \Lambda_i \le 2\alpha^2\sigma^2 + (6L_f^2 + 2\sigma^2)\Lambda_{i-1} + (4\alpha^2)\Gamma_i.$$

Remark: A key improvement over Lemma 3.4 is the reduction of the factor for Λ_{i-1} from $1/\alpha^2$ to $1/\alpha$ in the recurrence of Γ_i. This allows us to use a much smaller α to control the term $\alpha\sigma^2$, thereby removing the reliance on a large batch size B_1. We highlight this difference in the following lemma.

Lemma 4.2. *Let $D = 2(7L_f^2 + 2L_J C + 3\sigma^2)$. We obtain:*

$$\Gamma_i \le D^i \frac{1}{\alpha K}\sum_{k=1}^K \mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + (8D^i + 3)\alpha\sigma^2;$$

$$\Lambda_i \le D^i \frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + 8\alpha^2\sigma^2 D^i.$$

Remark: Compared to Lemma 3.5, the second term in the function estimation error changes from $\mathcal{O}(\sigma^2/B_1)$ to $\mathcal{O}(\alpha\sigma^2)$, which can be controlled by a small α rather than a large batch size. Also, the factor of the first term is improved from $1/\alpha^2$ to $1/\alpha$, enabling the use of much smaller α.

We then construct the overall gradient estimator using these $\mathbf{u}_i^k$ values without batches:

$$\mathbf{z}^k = (1-\alpha)\mathbf{z}^{k-1} + \alpha \prod_{i=1}^T \nabla f_i(\mathbf{u}_{i-1}^k; \xi_i^k), \tag{5}$$

where $\mathbf{z}^1 = \prod_{i=1}^T \nabla f_i(\mathbf{u}_{i-1}^1; \xi_i^1)$. Similarly, the gradient estimation error can be bounded as follows.

Lemma 4.3. *Let $D_0 = 2T^2(\sigma^2 + L_f^2)^T + 4TL_F^2\sigma^2(8D^T + 3T)$. We have the following guarantee:*

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\left\|\mathbf{z}^k - \nabla F(\mathbf{x}^k)\right\|^2\right] \le D_0\alpha + 4L_F^2\left(1 + TD^T\right)\frac{1}{\alpha^2 K}\sum_{k=1}^{K+1}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right].$$

Remark: By properly setting the learning rate η and momentum parameter α, we ensure a small gradient estimation error with a constant batch size, as $\frac{\sigma^2}{B_1} 4TL_F^2 \sum_{i=1}^{T-1} C_i$ in Lemma 3.6 is eliminated.

Finally, we update the decision variable $\mathbf{x}$ using an SGD-style step. The complete procedure, named Momentum-based Multi-Level method with First-order Approximation (MMLFA), is described in Algorithm 2. We now provide the convergence guarantee for MMLFA.

Theorem 4.4. *Under Assumptions 3.1, 3.2, 3.3 and supposing that* $\left\|f_i(\mathbf{u}_{i-1}^k)\right\| \leq C$*, by setting* $\eta = \mathcal{O}(K^{-1/2})$ *and* $\alpha = \mathcal{O}(K^{-1/2})$*, our MMLFA algorithm ensures that:*

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\left\|\nabla F(\mathbf{x}^k)\right\|\right] \leq \mathcal{O}\left(\frac{1}{K^{1/4}}\right).$$

Remark: To ensure $\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\left\|\nabla F(\mathbf{x}^k)\right\|\right] \leq \epsilon$, the sample complexity is $K = \mathcal{O}(\epsilon^{-4})$. This rate matches the previous result while eliminating the need for large batches.

5 Experiments

In this section, we validate the effectiveness of the proposed methods on mean-deviation risk-averse portfolio optimization and hierarchical tilted empirical risk minimization. We compare our methods with existing stochastic multi-level algorithms, including A-TSCGD (Yang et al., 2019), NLASG (Balasubramanian et al., 2022), Nested-SPIDER (Zhang & Xiao, 2021), and SCSC (Chen et al., 2021). Hyperparameters for each method are selected via grid search: the momentum parameter is tuned over the set $\{0.01, 0.05, 0.1, 0.5, 0.9, 0.95, 0.99\}$ and the learning rate is searched from the set $\{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. We report the best results obtained from the grid search for each method. If the optimal hyperparameter lies on the boundary of the grid, we extend the search range to ensure it lies within the interior. We set the batch size to 128 for each method. For the clipping threshold C, we observe that a sufficiently large value works well in practice, avoiding the need for tuning. All results are averaged over 20 runs.

5.1 Mean-deviation Risk-averse Portfolio Management

First, we investigate the risk-averse portfolio optimization. Consider a portfolio of d assets, where the vector $\mathbf{r}_n \in \mathbb{R}^d$ denotes the payoff of all assets at step $n \in \{1, \ldots, N\}$. We aim to identify a strategy $\mathbf{x}$, where $[\mathbf{x}]_i$ is the fraction of total capital allocated to asset i. The goal is to maximize expected return while controlling risk. A robust formulation for this problem is the mean-deviation risk-averse optimization model (Shapiro et al., 2021), in which risk is defined as the standard deviation. This model is widely applied in practice and frequently serves as a benchmark in multi-level optimization research (Yang et al., 2019; Zhang & Lan, 2021). The problem is formulated as:

$$\min_{\mathbf{x}} -\frac{1}{N}\sum_{n=1}^{N} \langle \mathbf{r}_n, \mathbf{x}\rangle + \lambda \sqrt{\frac{1}{N}\sum_{n=1}^{N} \left(\langle \mathbf{r}_n, \mathbf{x}\rangle - \langle \bar{\mathbf{r}}, \mathbf{x}\rangle\right)^2},$$

where $\bar{\mathbf{r}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{r}_n$. The above formulation constitutes a three-level stochastic compositional optimization problem, with levels defined as:

$$f_1(\mathbf{x}) = \left(\frac{1}{N}\sum_{n=1}^{N} \langle \mathbf{r}_n, \mathbf{x}\rangle, \mathbf{x}\right),$$

$$f_2((z, \mathbf{x})) = \left(z, \frac{1}{N}\sum_{n=1}^{N} \left(\langle \mathbf{r}_n, \mathbf{x}\rangle - z\right)^2\right),$$

$$f_3\left((y_1, y_2)\right) = -y_1 + \lambda\sqrt{y_2}.$$

In the experiments, we evaluate on three real-world datasets (Industry-10, Industry-12, and Industry-17) from the Kenneth R. French Data Library[1], which contain 10, 12, and 17 industrial assets,

[1] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/

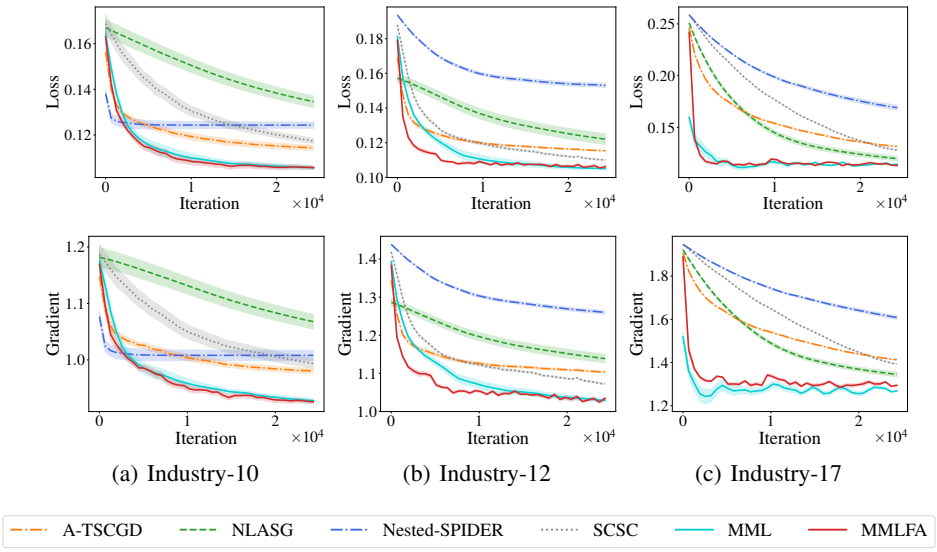


Figure 1: Results for Risk-Averse Portfolio Optimization.

respectively. Following Zhang & Xiao (2021), we set $\lambda = 0.2$. Figure 1 illustrates the loss values and gradient norms with respect to iterations (since the batch size is fixed to 128, the sample complexity is just proportional to the iteration number). Across all datasets, our methods (MML and MMLFA) demonstrate superior performance. Both the loss and gradient norms decrease more rapidly and stabilize at lower levels compared to the other methods. MML and MMLFA perform similarly overall: MMLFA performs slightly better on Industry-10/12, while MML is slightly better on Industry-17.

Note that although MML requires large batch sizes in the theoretical analysis, we do not observe significant performance degradation in the experiments when using smaller batches. In Appendix K, we validate the performance with different batch sizes, i.e., $\{48, 64, 96, 128\}$. As shown in Figure 3, we do not observe significant performance degradation with smaller batches, demonstrating the practical robustness of MML with small batches. Overall, we would recommend using MML first, since it is simpler and therefore easier and more stable to implement. When the batch size is extremely small, MMLFA may be preferable.

We further study the effect of the clipping threshold C in MMLFA. Theoretically, choosing C above the true upper bound of the relevant function values does not change the order of the convergence rate. Here, we vary C within the range $\{0.5, 1, 5, 10, 10^{20}\}$, with results depicted in Figure 4 in Appendix K. The results indicate that performance is stable provided C is not too small (e.g., $C \geq 1$). In contrast, overly aggressive clipping (e.g., $C = 0.5$) degrades performance, suggesting that a large default value of C is reasonable in practice.

5.2 Hierarchical Tilted Empirical Risk Minimization

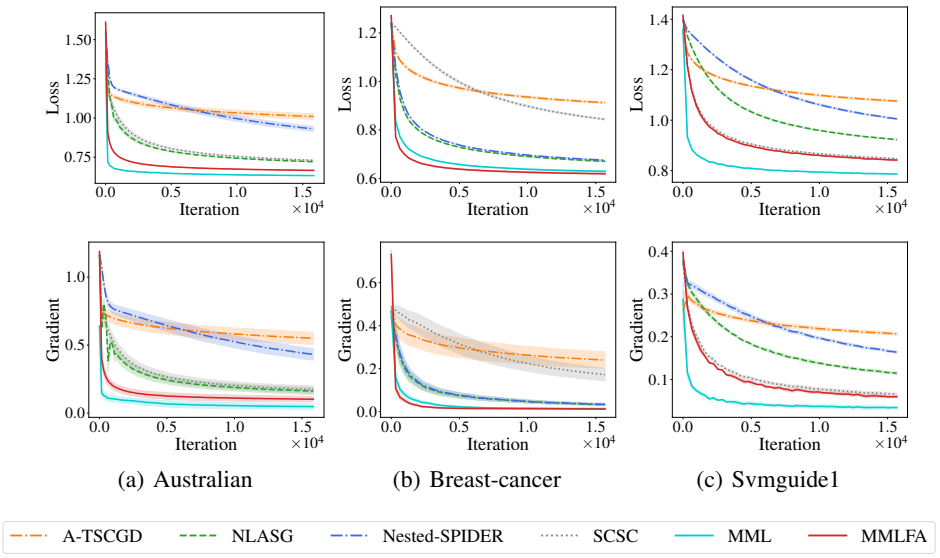
The Hierarchical Tilted Empirical Risk Minimization (TERM) objective (Li et al., 2021; 2023) is a loss function designed to simultaneously achieve robustness against noisy data and ensure fairness across subgroups. Given samples $\{z_1, \ldots, z_N\}$ and per-sample loss $f(\theta; z_i)$, the TERM objective is defined as $\widetilde{R}(\theta) := \frac{1}{t} \log \left(\frac{1}{N} \sum_{i \in [N]} e^{t f(\theta; z_i)} \right)$. It mitigates outliers when $t < 0$ and can help address class imbalance when $t > 0$. When a task involves both outliers and class imbalance, the Hierarchical TERM approach can be employed:

$$\tilde{J}(w) := \frac{1}{t} \log \left(\frac{1}{|\mathcal{D}|} \sum_{\mathcal{G} \subseteq \mathcal{D}} |\mathcal{G}| e^{t \tilde{R}_{\mathcal{G}}(w)} \right), \text{ with } \widetilde{R}_{\mathcal{G}}(w) := \frac{1}{\tau} \log \left(\frac{1}{|\mathcal{G}|} \sum_{z \in \mathcal{G}} e^{\tau \ell(w; z)} \right),$$

where $\mathcal{D}$ represents the set of all training samples and $\mathcal{G}$ denotes the samples for a specific class. The parameters t and τ are constants addressing outliers and class imbalance, respectively. This

Table 1: Classification accuracy (mean±std) for Hierarchical Tilted Empirical Risk Minimization.

Method	Australian		Breast-cancer		Svmguide1	
	rare class	overall	rare class	overall	rare class	overall
A-TSCGD	0.609±0.011	0.717±0.007	0.610±0.007	0.819±0.001	0.744±0.002	0.861±0.001
SCSC	0.723±0.004	0.782±0.002	0.668±0.009	0.841±0.004	0.777±0.008	0.881±0.003
Nested-SPIDER	0.692±0.010	0.754±0.008	0.747±0.010	0.877±0.005	0.762±0.003	0.864±0.005
NLASG	0.750±0.006	0.790±0.005	0.728±0.011	0.870±0.006	0.754±0.002	0.866±0.002
MML	0.813±0.004	**0.823±0.006**	0.805±0.004	0.913±0.003	**0.782±0.002**	**0.884±0.002**
MMLFA	**0.838±0.007**	0.817±0.004	**0.824±0.002**	**0.920±0.002**	0.778±0.001	0.882±0.001



(a) Australian (b) Breast-cancer (c) Svmguide1

Figure 2: Results for Hierarchical Tilted Empirical Risk Minimization.

formulation yields a four-level stochastic compositional optimization problem:

$$f_1(w) = \frac{1}{|\mathcal{G}|}\sum_{z\in\mathcal{G}} e^{\tau\ell(w;z)},\ f_2(x) = \frac{1}{\tau}\log(x), f_3(y) = \frac{1}{|\mathcal{D}|}\sum_{\mathcal{G}\subseteq\mathcal{D}} |\mathcal{G}|e^{ty},\ f_4(z) = \frac{1}{t}\log(z).$$

In the experiments, we evaluate the different methods using the `australian`, `breast-cancer`, and `svmguide1` datasets from LIBSVM (Chang & Lin, 2011). We inject both label noise and class imbalance: approximately 30% of labels are randomly reshuffled, and the rare-to-common class ratio is set to 1:20. As shown in Figure 2, MML and MMLFA decrease both the objective and gradient norm faster than the baselines. Table 1 reports classification accuracy (mean±std over 20 runs), showing that our methods achieve the best performance on both the rare class and overall accuracy. Finally, Figure 5 in Appendix K provides a sensitivity analysis for the clipping threshold C, indicating that performance is robust when C is sufficiently large (e.g., $C > 1$).

6 Conclusion

In this paper, we addressed stochastic multi-level optimization by employing momentum-based estimators to track function values and gradients. With mini-batches, the resulting method achieves the optimal sample complexity $\mathcal{O}(\epsilon^{-4})$ under standard smoothness assumptions, without requiring the stronger average smoothness conditions used in prior work. We further introduced a normalized variant that preserves the same rate while avoiding dependence on problem-specific constants to set hyperparameters. To eliminate large batches, we developed a batch-free approach based on first-order approximation and clipping, which retains the optimal sample complexity with a constant batch size. Experiments on risk-averse portfolio optimization and hierarchical tilted empirical risk minimization further demonstrate the effectiveness of the proposed methods.

AI USE STATEMENT

In this work, we used generative AI tools mainly for polishing writing. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake E. Woodworth. Lower bounds for non-convex stochastic optimization. *ArXiv e-prints*, arXiv:1912.02365, 2019.

Krishnakumar Balasubramanian, Saeed Ghadimi, and Anthony Nguyen. Stochastic multilevel composition optimization algorithms with level-independent convergence rates. *SIAM Journal on Optimization*, 32(2):519–544, 2022.

Sergio Bruno, Shabbir Ahmed, Alexander Shapiro, and Alexandre Street. Risk neutral and risk averse approaches to multistage renewable investment planning under uncertainty. *European Journal of Operational Research*, 250(3):979–989, 2016.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, 69:4937–4948, 2021.

Shawn Cole, Xavier Giné, and James Vickery. How Does Risk Management Influence Production Decisions? Evidence from a Field Experiment. *The Review of Financial Studies*, 30(6):1935–1970, 2017.

Weilin Cong, Rana Forsati, Mahmut Kandemir, and Mehrdad Mahdavi. Minimal variance sampling with provable guarantees for fast training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1393–1403, 2020.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems 32*, pp. 15210–15219, 2019.

Christoph Dann, Gerhard Neumann, and Jan Peters. Policy evaluation with temporal differences: a survey and comparison. *Journal of Machine Learning Research*, 15:809–883, 2014.

Darinka Dentcheva, Spiridon I. Penev, and A. Ruszczynski. Statistical estimation of composite risk functionals and risk optimization problems. *Annals of the Institute of Statistical Mathematics*, 69 (4):737–760, 2017.

Hongchang Gao. A doubly recursive stochastic compositional gradient descent method for federated multi-level compositional optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 14540–14610, 2024a.

Hongchang Gao. Decentralized multi-level compositional optimization algorithms with level-independent convergence rate. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pp. 4402–4410, 2024b.

Florian Hübler, Junchi Yang, Xiang Li, and Niao He. Parameter-agnostic optimization under relaxed smoothness. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pp. 4861–4869, 2024.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Multi-step model-agnostic meta-learning: Convergence and improved algorithms. *ArXiv e-prints*, arXiv:2002.07836, 2020.

Wei Jiang, Bokun Wang, Yibo Wang, Lijun Zhang, and Tianbao Yang. Optimal algorithms for stochastic multi-level compositional optimization. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 10195–10216, 2022.

Wei Jiang, Sifan Yang, Wenhao Yang, Yibo Wang, Yuanyu Wan, and Lijun Zhang. Projection-free variance reduction methods for stochastic constrained multi-level compositional optimization. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 21962–21987, 2024.

Wei Jiang, Sifan Yang, Yibo Wang, Tianbao Yang, and Lijun Zhang. Revisiting stochastic multi-level compositional optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47 (7):5613–5624, 2025.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. In *International Conference on Learning Representations*, 2021.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. On tilted losses in machine learning: Theory and applications. *Journal of Machine Learning Research*, 24(142):1–79, 2023.

Xiaokang Pan, Xingyu Li, Jin Liu, Tao Sun, Kai Sun, Lixing Chen, and Zhe Qu. Stability and generalization for stochastic recursive momentum-based algorithms for (Strongly-)Convex one to k-level stochastic optimizations. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 39201–39275, 2024.

Andrzej Ruszczyński. A stochastic subgradient method for nonsmooth nonconvex multilevel composition optimization. *SIAM Journal on Control and Optimization*, 59(3):2301–2320, 2021.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. Society for Industrial and Applied Mathematics, 2021.

Tesi Xiao, Krishnakumar Balasubramanian, and Saeed Ghadimi. A projection-free algorithm for constrained stochastic multi-level composition optimization. In *Advances in Neural Information Processing Systems 35*, pp. 19984–19996, 2022.

Shuoguang Yang and Fengpei Li. A communication-efficient algorithm for federated multilevel stochastic compositional optimization. *IEEE Transactions on Signal Processing*, 72:2333–2347, 2024.

Shuoguang Yang, Mengdi Wang, and Ethan X. Fang. Multilevel stochastic gradient methods for nested composition optimization. *SIAM Journal on Optimization*, 29(1):616–659, 2019.

Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9):2805–2824, 2019.

Junyu Zhang and Lin Xiao. Multilevel composite stochastic optimization via nested variance reduction. *SIAM Journal on Optimization*, 31(2):1131–1157, 2021.

Xinwen Zhang and Hongchang Gao. On the convergence of stochastic smoothed multi-level compositional gradient descent ascent. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Xinwen Zhang, Ali Payani, Myungjin Lee, Richard Souvenir, and Hongchang Gao. A federated stochastic multi-level compositional minimax algorithm for deep AUC maximization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 59601–59640, 2024.

Zhe Zhang and Guanghui Lan. Optimal algorithms for convex nested stochastic composite optimization. *ArXiv e-prints*, arXiv:2011.10076, 2021.

Zhe Zhang and Guanghui Lan. Optimal methods for convex nested stochastic composite optimization. *Mathematical Programming*, 212(1):1–48, 2025.

A PROOF OF LEMMA 3.4

Considering that $\mathbf{u}_i^k = (1-\alpha)\mathbf{u}_i^{k-1} + \alpha\frac{1}{B_1}\sum_{j=1}^{B_1} f_i(\mathbf{u}_{i-1}^k;\xi_{i,j}^k)$, we have:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{u}_i^k - f_i(\mathbf{u}_{i-1}^k)\right\|^2\right] \\
=&\mathbb{E}\left[\left\|(1-\alpha)\left(\mathbf{u}_i^{k-1} - f_i(\mathbf{u}_{i-1}^k)\right) + \alpha\left(\frac{1}{B_1}\sum_{j=1}^{B_1} f_i(\mathbf{u}_{i-1}^k;\xi_{i,j}^k) - f_i(\mathbf{u}_{i-1}^k)\right)\right\|^2\right] \\
=&(1-\alpha)^2\mathbb{E}\left[\left\|\mathbf{u}_i^{k-1} - f_i(\mathbf{u}_{i-1}^k)\right\|^2\right] + \alpha^2\mathbb{E}\left[\left\|\frac{1}{B_1}\sum_{j=1}^{B_1} f_i(\mathbf{u}_{i-1}^k;\xi_{i,j}^k) - f_i(\mathbf{u}_{i-1}^k)\right\|^2\right] \\
\leq&(1-\alpha)^2(1+\alpha)\mathbb{E}\left[\left\|\mathbf{u}_i^{k-1} - f_i(\mathbf{u}_{i-1}^{k-1})\right\|^2\right] + (1-\alpha)^2(1+\frac{1}{\alpha})\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^{k-1}) - f_i(\mathbf{u}_{i-1}^k)\right\|^2\right] + \frac{\alpha^2\sigma^2}{B_1} \\
\leq&(1-\alpha)\mathbb{E}\left[\left\|\mathbf{u}_i^{k-1} - f_i(\mathbf{u}_{i-1}^{k-1})\right\|^2\right] + \frac{2}{\alpha}\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^{k-1}) - f_i(\mathbf{u}_{i-1}^k)\right\|^2\right] + \frac{\alpha^2\sigma^2}{B_1} \\
\leq&(1-\alpha)\mathbb{E}\left[\left\|\mathbf{u}_i^{k-1} - f_i(\mathbf{u}_{i-1}^{k-1})\right\|^2\right] + \frac{2L_f^2}{\alpha}\mathbb{E}\left[\left\|\mathbf{u}_{i-1}^{k-1} - \mathbf{u}_{i-1}^k\right\|^2\right] + \frac{\alpha^2\sigma^2}{B_1}.
\end{aligned}
$$

Set $B_1 \leq \alpha^2 B_0 K$. Summing up and rearranging, we can obtain:

$$
\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{u}_i^k - f_i(\mathbf{u}_{i-1}^k)\right\|^2\right] \leq&\frac{\mathbb{E}\left[\left\|\mathbf{u}_i^1 - f_i(\mathbf{u}_{i-1}^1)\right\|^2\right]}{\alpha K} + \frac{2L_f^2}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{u}_{i-1}^{k+1} - \mathbf{u}_{i-1}^k\right\|^2\right] + \frac{\alpha\sigma^2}{B_1} \\
\leq&\frac{\sigma^2}{\alpha K B_0} + \frac{2L_f^2}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{u}_{i-1}^{k+1} - \mathbf{u}_{i-1}^k\right\|^2\right] + \frac{\alpha\sigma^2}{B_1} \\
\leq&\frac{2L_f^2}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{u}_{i-1}^{k+1} - \mathbf{u}_{i-1}^k\right\|^2\right] + \frac{2\alpha\sigma^2}{B_1}.
\end{aligned}
$$

Next, we bound the difference of function estimators between adjacent steps:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{u}_i^{k+1} - \mathbf{u}_i^k\right\|^2\right] = \alpha^2\mathbb{E}\left[\left\|\frac{1}{B_1}\sum_{j=1}^{B_1} f_i(\mathbf{u}_{i-1}^{k+1};\xi_{i,j}^{k+1}) - \mathbf{u}_i^k\right\|^2\right] \\
=&\alpha^2\mathbb{E}\left[\left\|\frac{1}{B_1}\sum_{j=1}^{B_1} f_i(\mathbf{u}_{i-1}^{k+1};\xi_{i,j}^{k+1}) - f_i(\mathbf{u}_{i-1}^{k+1}) + f_i(\mathbf{u}_{i-1}^{k+1}) - \mathbf{u}_i^k\right\|^2\right] \\
=&\alpha^2\mathbb{E}\left[\left\|\frac{1}{B_1}\sum_{j=1}^{B_1} f_i(\mathbf{u}_{i-1}^{k+1};\xi_{i,j}^{k+1}) - f_i(\mathbf{u}_{i-1}^{k+1})\right\|^2\right] + \alpha^2\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^{k+1}) - \mathbf{u}_i^k\right\|^2\right] \\
\leq&\frac{\alpha^2\sigma^2}{B_1} + \alpha^2\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^{k+1}) - f_i(\mathbf{u}_{i-1}^k) + f_i(\mathbf{u}_{i-1}^k) - \mathbf{u}_i^k\right\|^2\right] \\
\leq&\frac{\alpha^2\sigma^2}{B_1} + 2\alpha^2\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^{k+1}) - f_i(\mathbf{u}_{i-1}^k)\right\|^2\right] + 2\alpha^2\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^k) - \mathbf{u}_i^k\right\|^2\right] \\
\leq&\frac{\alpha^2\sigma^2}{B_1} + 2\alpha^2 L_f^2\mathbb{E}\left[\left\|\mathbf{u}_{i-1}^{k+1} - \mathbf{u}_{i-1}^k\right\|^2\right] + 2\alpha^2\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^k) - \mathbf{u}_i^k\right\|^2\right].
\end{aligned}
$$

Summing up, we obtain that

$$
\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{u}_i^{k+1} - \mathbf{u}_i^k\right\|^2\right] \leq \frac{\alpha^2\sigma^2}{B_1} + \frac{2\alpha^2 L_f^2}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{u}_{i-1}^{k+1} - \mathbf{u}_{i-1}^k\right\|^2\right] + \frac{2\alpha^2}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^k) - \mathbf{u}_i^k\right\|^2\right].
$$

B PROOF OF LEMMA 3.5

We would prove this Lemma by Induction.

(1) For level $i = 1$, by noting that $\mathbf{u}_0^k = \mathbf{x}^k$ and setting $C_1 = 5 + 6L_f^2$, we have

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{u}_1^k - f_1(\mathbf{u}_0^k)\right\|^2\right] \\
\leq&\frac{2L_f^2}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{u}_0^{k+1} - \mathbf{u}_0^k\right\|^2\right] + \frac{2\alpha\sigma^2}{B_1} \leq \frac{2L_f^2}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + \frac{2\alpha\sigma^2}{B_1} \\
\leq&(2L_f^2 + 2)\left(\frac{1}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + \frac{\sigma^2}{B_1}\right) \\
\leq&C_1\left(\frac{1}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + \frac{\sigma^2}{B_1}\right),
\end{aligned}$$

as well as the following guarantee

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{u}_1^{k+1} - \mathbf{u}_1^k\right\|^2\right] \\
\leq&\frac{\alpha^2\sigma^2}{B_1} + \frac{2\alpha^2 L_f^2}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{u}_0^{k+1} - \mathbf{u}_0^k\right\|^2\right] + \frac{2\alpha^2}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|f_1(\mathbf{u}_0^k) - \mathbf{u}_1^k\right\|^2\right] \\
\leq&\frac{\alpha^2\sigma^2}{B_1} + 2\alpha^2 L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + 2\alpha^2\left(\frac{2L_f^2}{\alpha^2}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + \frac{2\alpha\sigma^2}{B_1}\right) \\
\leq&(6L_f^2 + 5)\alpha^2\left(\frac{1}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + \frac{\sigma^2}{B_1}\right) \\
\leq&C_1\alpha^2\left(\frac{1}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + \frac{\sigma^2}{B_1}\right).
\end{aligned}$$

(2) Suppose that for level $i = n$, we have that

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{u}_n^k - f_n(\mathbf{u}_{n-1}^k)\right\|^2\right] \leq C_n\left(\frac{1}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + \frac{\sigma^2}{B_1}\right), \\
\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{u}_n^{k+1} - \mathbf{u}_n^k\right\|^2\right] \leq C_n\alpha^2\left(\frac{1}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + \frac{\sigma^2}{B_1}\right).
\end{aligned}$$

Then, for level $i = n + 1$, we can prove that:

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{u}_{n+1}^k - f_{n+1}(\mathbf{u}_n^k)\right\|^2\right] \leq&\frac{2L_f^2}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{u}_n^{k+1} - \mathbf{u}_n^k\right\|^2\right] + \frac{2\alpha\sigma^2}{B_1} \\
\leq&\frac{2L_f^2}{\alpha^2}C_n\alpha^2\left(\frac{1}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + \frac{\sigma^2}{B_1}\right) + \frac{2\alpha\sigma^2}{B_1} \\
\leq&(2 + 2L_f^2 C_n)\left(\frac{1}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + \frac{\sigma^2}{B_1}\right) \\
\leq&C_{n+1}\left(\frac{1}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + \frac{\sigma^2}{B_1}\right),
\end{aligned}$$

as well as the fact that

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{u}_{n+1}^{k+1}-\mathbf{u}_{n+1}^{k}\right\|^2\right] \leq & \frac{\alpha^2\sigma^2}{B_1}+\frac{2\alpha^2L_f^2}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{u}_{n}^{k+1}-\mathbf{u}_{n}^{k}\right\|^2\right]+\frac{2\alpha^2}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{u}_{n+1}^{k}-f_{n+1}(\mathbf{u}_n^k)\right\|^2\right] \\
\leq & \frac{\alpha^2\sigma^2}{B_1}+2\alpha^2L_f^2C_n\alpha^2\left(\frac{1}{\alpha^2K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\right\|^2\right]+\frac{\sigma^2}{B_1}\right) \\
& +2\alpha^2(2+2L_f^2C_n)\left(\frac{1}{\alpha^2K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\right\|^2\right]+\frac{\sigma^2}{B_1}\right) \\
\leq & (5+6L_f^2C_n)\alpha^2\left(\frac{1}{\alpha^2K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\right\|^2\right]+\frac{\sigma^2}{B_1}\right) \\
\leq & C_{n+1}\alpha^2\left(\frac{1}{\alpha^2K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\right\|^2\right]+\frac{\sigma^2}{B_1}\right).
\end{aligned}$$

Thus, we have proved this lemma via Induction. Also, by setting that $C_0 = 1$, we have $C_{i+1} = 5 + 6L_f^2C_i$ for $i \geq 1$.

C Proof of Lemma 3.6

First, we define $\mathbf{y}_i^k = f_i \circ f_{i-1} \circ \ldots \circ f_1(\mathbf{x}^k)$, and define its gradient as $\nabla F_i(\mathbf{x}^k)$. Also denote that $\nabla \widehat{F}_i(\mathbf{x}^k) \coloneqq \nabla f_1(\mathbf{x}^k) \cdots \nabla f_i(\mathbf{u}_{i-1}^k)$. Then, we have the following guarantee:

$$\begin{aligned}
\left\|\mathbf{y}_1^k-\mathbf{u}_1^k\right\| &= \left\|f_1(\mathbf{x}^k)-\mathbf{u}_1^k\right\| \\
\left\|\mathbf{y}_2^k-\mathbf{u}_2^k\right\| &= \left\|f_2\circ f_1(\mathbf{x}^k)-\mathbf{u}_2^k\right\| \\
&\leq \left\|f_2\circ f_1(\mathbf{x}^k)-f_2(\mathbf{u}_1^k)\right\|+\left\|f_2(\mathbf{u}_1^k)-\mathbf{u}_2^k\right\| \\
&\leq L_f\left\|f_1(\mathbf{x}^k)-\mathbf{u}_1^k\right\|+\left\|f_2(\mathbf{u}_1^k)-\mathbf{u}_2^k\right\|, \\
\left\|\mathbf{y}_3^k-\mathbf{u}_3^k\right\| &= \left\|f_3\circ f_2\circ f_1(\mathbf{x}^k)-\mathbf{u}_3^k\right\| \\
&\leq \left\|f_3\circ f_2\circ f_1(\mathbf{x}^k)-f_3(\mathbf{u}_2^k)\right\|+\left\|f_3(\mathbf{u}_2^k)-\mathbf{u}_3^k\right\| \\
&\leq L_f\left\|f_2\circ f_1(\mathbf{x}^k)-\mathbf{u}_2^k\right\|+\left\|f_3(\mathbf{u}_2^k)-\mathbf{u}_3^k\right\| \\
&\leq L_f\left\|\mathbf{y}_2^k-\mathbf{u}_2^k\right\|+\left\|f_3(\mathbf{u}_2^k)-\mathbf{u}_3^k\right\| \\
&\leq L_f(L_f\left\|f_1(\mathbf{x}^k)-\mathbf{u}_1^k\right\|+\left\|f_2(\mathbf{u}_1^k)-\mathbf{u}_2^k\right\|)+\left\|f_3(\mathbf{u}_2^k)-\mathbf{u}_3^k\right\| \\
&\cdots \\
\left\|\mathbf{y}_i^k-\mathbf{u}_i^k\right\| &\leq L_f\left\|\mathbf{y}_{i-1}^k-\mathbf{u}_{i-1}^k\right\|+\left\|f_i(\mathbf{u}_{i-1}^k)-\mathbf{u}_i^k\right\| \\
&\leq L_f\left(L_f\left\|\mathbf{y}_{i-2}^k-\mathbf{u}_{i-2}^k\right\|+\left\|f_{i-1}(\mathbf{u}_{i-2}^k)-\mathbf{u}_{i-1}^k\right\|\right)+\left\|f_i(\mathbf{u}_{i-1}^k)-\mathbf{u}_i^k\right\| \\
&\leq \sum_{j=1}^{i}L_f^{i-j}\left\|f_j(\mathbf{u}_{j-1}^k)-\mathbf{u}_j^k\right\|.
\end{aligned}$$

Besides, we also know that:

$$\begin{aligned}
&\left\|\nabla F_1(\mathbf{x}^k)-\nabla\widehat{F}_1(\mathbf{x}^k)\right\| = 0, \\
&\left\|\nabla F_2(\mathbf{x}^k)-\nabla\widehat{F}_2(\mathbf{x}^k)\right\| = \left\|\nabla f_1(\mathbf{x}^k)\nabla f_2(\mathbf{y}_1^k)-\nabla f_1(\mathbf{x}^k)\nabla f_2(\mathbf{u}_1^k)\right\| \leq L_fL_J\left\|\mathbf{y}_1^k-\mathbf{u}_1^k\right\|, \\
&\left\|\nabla F_3(\mathbf{x}^k)-\nabla\widehat{F}_3(\mathbf{x}^k)\right\| = \left\|\nabla f_1(\mathbf{x}^k)\nabla f_2(\mathbf{y}_1^k)\nabla f_3(\mathbf{y}_2^k)-\nabla f_1(\mathbf{x}^k)\nabla f_2(\mathbf{u}_1^k)\nabla f_3(\mathbf{u}_2^k)\right\| \\
&\qquad\qquad \leq L_f\left\|\nabla f_2(\mathbf{y}_1^k)\nabla f_3(\mathbf{y}_2^k)-\nabla f_2(\mathbf{u}_1^k)\nabla f_3(\mathbf{u}_2^k)\right\| \\
&\qquad\qquad = L_f\left\|\nabla f_2(\mathbf{y}_1^k)\left(\nabla f_3(\mathbf{y}_2^k)-\nabla f_3(\mathbf{u}_2^k)\right)+\left(\nabla f_2(\mathbf{y}_1^k)-\nabla f_2(\mathbf{u}_1^k)\right)\nabla f_3(\mathbf{u}_2^k)\right\| \\
&\qquad\qquad \leq L_f^2L_J\left(\left\|\mathbf{y}_2^k-\mathbf{u}_2^k\right\|+\left\|\mathbf{y}_1^k-\mathbf{u}_1^k\right\|\right), \\
&\qquad\qquad \cdots
\end{aligned}$$

To this end, we can conclude that:

$$\left\|\nabla F_T(\mathbf{x}^k) - \nabla \widehat{F}_T(\mathbf{x}^k)\right\| \leq L_f^{T-1} L_J \sum_{i=1}^{T-1} \left\|\mathbf{y}_i^k - \mathbf{u}_i^k\right\|$$
$$\leq L_f^{T-1} L_J (1 + L_f + \ldots + L_f^{T-1}) \sum_{i=1}^{T-1} \left\|f_i(\mathbf{u}_{i-1}^k) - \mathbf{u}_i^k\right\| \leq L_F \sum_{i=1}^{T-1} \left\|f_i(\mathbf{u}_{i-1}^k) - \mathbf{u}_i^k\right\|,$$

and thus we can obtain

$$\left\|\nabla F(\mathbf{x}^k) - \prod_{i=1}^{T} \nabla f_i(\mathbf{u}_{i-1}^k)\right\|^2 \leq L_F^2 T \sum_{i=1}^{T-1} \left\|f_i(\mathbf{u}_{i-1}^k) - \mathbf{u}_i^k\right\|^2.$$

Next, considering the definition of $\mathbf{z}^{k+1}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{z}^{k+1} - \nabla F\left(\mathbf{x}^{k+1}\right)\right\|^2\right] \\
=&\mathbb{E}\left[\left\|(1-\alpha)(\mathbf{z}^k - \nabla F(\mathbf{x}^{k+1})) + \alpha\left(\prod_{i=1}^{T} \nabla f_i(\mathbf{u}_{i-1}^{k+1}) - \nabla F(\mathbf{x}^{k+1})\right)\right.\right. \\
&\left.\left. \quad + \alpha\left(\frac{1}{B_1}\sum_{j=1}^{B_1}\left[\prod_{i=1}^{T} \nabla f_i(\mathbf{u}_{i-1}^{k+1}; \xi_{i,j}^{k+1})\right] - \prod_{i=1}^{T} \nabla f_i(\mathbf{u}_{i-1}^{k+1})\right)\right\|^2\right] \\
\leq&\mathbb{E}\left[\left\|(1-\alpha)(\mathbf{z}^k - \nabla F(\mathbf{x}^{k+1})) + \alpha\left(\prod_{i=1}^{T} \nabla f_i(\mathbf{u}_{i-1}^{k+1}) - \nabla F(\mathbf{x}^{k+1})\right)\right\|^2\right] \\
&+ \alpha^2 \mathbb{E}\left[\left\|\frac{1}{B_1}\sum_{j=1}^{B_1}\left[\prod_{i=1}^{T} \nabla f_i(\mathbf{u}_{i-1}^{k+1}; \xi_{i,j}^{k+1})\right] - \prod_{i=1}^{T} \nabla f_i(\mathbf{u}_{i-1}^{k+1})\right\|^2\right] \\
\leq&\mathbb{E}\left[\left\|(1-\alpha)(\mathbf{z}^k - \nabla F(\mathbf{x}^k)) + (1-\alpha)(\nabla F(\mathbf{x}^k) - \nabla F(\mathbf{x}^{k+1})) + \alpha\left(\prod_{i=1}^{T} \nabla f_i(\mathbf{u}_{i-1}^{k+1}) - \nabla F(\mathbf{x}^{k+1})\right)\right\|^2\right] \\
&+ \frac{\alpha^2}{B_1} T^2 (\sigma^2 + L_f^2)^{T-1} \sigma^2 \\
\leq&(1-\alpha)\mathbb{E}\left[\left\|\mathbf{z}^k - \nabla F(\mathbf{x}^k)\right\|^2\right] + \frac{4}{\alpha}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}^k) - \nabla F(\mathbf{x}^{k+1})\right\|^2\right] + 4\alpha\mathbb{E}\left[\left\|\prod_{i=1}^{T} \nabla f_i(\mathbf{u}_{i-1}^{k+1}) - \nabla F(\mathbf{x}^{k+1})\right\|^2\right] \\
&+ \frac{\alpha^2}{B_1} T^2 (\sigma^2 + L_f^2)^{T-1} \sigma^2 \\
\leq&(1-\alpha)\mathbb{E}\left[\left\|\mathbf{z}^k - \nabla F(\mathbf{x}^k)\right\|^2\right] + \frac{4L_F^2}{\alpha}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + 4\alpha T L_F^2 \sum_{i=1}^{T-1} \mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^{k+1}) - \mathbf{u}_i^{k+1}\right\|^2\right] \\
&+ \frac{\alpha^2}{B_1} T^2 (\sigma^2 + L_f^2)^T.
\end{aligned}$$

Summing up and rearranging, we proved the conclusion by noting that $B_1 \leq \alpha^2 B_0 K$ and

$$\mathbb{E}\left[\left\|\mathbf{z}^1 - \nabla F(\mathbf{x}^1)\right\|^2\right] \leq T^2(\sigma^2 + L_f^2)^T / B_0 \leq \alpha^2 K T^2 (\sigma^2 + L_f^2)^T / B_1,$$
$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\left\|\mathbf{u}_i^{k+1} - f_i(\mathbf{u}_{i-1}^{k+1})\right\|^2\right] \leq \frac{1}{K}\sum_{k=1}^{K+1} \mathbb{E}\left[\left\|\mathbf{u}_i^k - f_i(\mathbf{u}_{i-1}^k)\right\|^2\right] \leq C_i\left(\frac{1}{\alpha^2 K}\sum_{k=1}^{K+1} \mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + \frac{\sigma^2}{B_1}\right).$$

D PROOF OF THEOREM 3.7

Due to the smoothness of function F and the definition of $\mathbf{x}^{k+1}$, by setting $\eta \leq \frac{1}{2L_F}$, we have:

$$
\begin{aligned}
F\left(\mathbf{x}^{k+1}\right) &\leq F\left(\mathbf{x}^{k}\right) + \left\langle \nabla F(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \right\rangle + \frac{L_F}{2}\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2 \\
&= F(\mathbf{x}^k) - \eta \left\langle \nabla F(\mathbf{x}^k), \mathbf{z}^k \right\rangle + \frac{\eta^2 L_F}{2}\|\mathbf{z}^k\|^2 \\
&\leq F(\mathbf{x}^k) - \eta\langle \nabla F(\mathbf{x}^k), \mathbf{z}^k\rangle + \frac{\eta}{2}\|\nabla F(\mathbf{x}^k)\|^2 + \frac{\eta}{2}\left\|\mathbf{z}^k\right\|^2 - \frac{\eta}{2}\|\nabla F(\mathbf{x}^k)\|^2 - \frac{\eta}{2}\left\|\mathbf{z}^k\right\|^2 + \frac{\eta}{4}\left\|\mathbf{z}^k\right\|^2 \\
&= F(\mathbf{x}^k) + \frac{\eta}{2}\|\nabla F(\mathbf{x}^k) - \mathbf{z}^k\|^2 - \frac{\eta}{2}\|\nabla F(\mathbf{x}^k)\|^2 - \frac{\eta}{4}\left\|\mathbf{z}^k\right\|^2.
\end{aligned}
$$

Also note that $\mathbb{E}\left[\left\|\mathbf{z}^1\right\|^2\right] \leq 2(L_f^2 + \sigma^2)^T$ and

$$
\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{z}^k\right\|^2\right] =& \mathbb{E}\left[\left\|(1-\alpha)\mathbf{z}^{k-1} + \alpha\frac{1}{B_1}\sum_{j=1}^{B_1}\prod_{i=1}^{T}\nabla f_i(\mathbf{u}_{i-1}^k;\xi_{i,j}^k)\right\|^2\right] \\
\leq& (1-\alpha)^2(1+\alpha)\mathbb{E}\left[\left\|\mathbf{z}^{k-1}\right\|^2\right] + \alpha^2(1+\frac{1}{\alpha})\mathbb{E}\left[\left\|\frac{1}{B_1}\sum_{j=1}^{B_1}\prod_{i=1}^{T}\nabla f_i(\mathbf{u}_{i-1}^k;\xi_{i,j}^k)\right\|^2\right] \\
\leq& (1-\alpha)\mathbb{E}\left[\left\|\mathbf{z}^{k-1}\right\|^2\right] + 2\alpha\frac{1}{B_1}\sum_{j=1}^{B_1}\mathbb{E}\left[\left\|\prod_{i=1}^{T}\nabla f_i(\mathbf{u}_{i-1}^k;\xi_{i,j}^k)\right\|^2\right] \\
\leq& (1-\alpha)\mathbb{E}\left[\left\|\mathbf{z}^{k-1}\right\|^2\right] + 2\alpha\frac{1}{B_1}\sum_{j=1}^{B_1}\prod_{i=1}^{T}\mathbb{E}_{\xi_{i,j}}\left[\left\|\nabla f_i(\mathbf{u}_{i-1}^k;\xi_{i,j}^k)\right\|^2\right] \\
\leq& (1-\alpha)\mathbb{E}\left[\left\|\mathbf{z}^{k-1}\right\|^2\right] + \alpha\cdot 2\left(L_f^2+\sigma^2\right)^T \leq 2\left(L_f^2+\sigma^2\right)^T.
\end{aligned}
$$

Set $B_0 = 1$, $\alpha = \mathcal{O}(1)$, $B_1 = \alpha^2 K = \mathcal{O}(K)$ and learning rate $\eta \leq \frac{\alpha}{4L_F\sqrt{1+T\sum_{i=1}^{T-1}C_i}} = \mathcal{O}(1)$. We can ensure

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla F(\mathbf{x}^k)\right\|^2\right] \\
\leq& \frac{2\mathbb{E}\left[F(\mathbf{x}^1) - F(\mathbf{x}^{K+1})\right]}{\eta K} + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}^k) - \mathbf{z}^k\right\|^2\right] - \frac{1}{2K}\mathbb{E}\left[\sum_{k=1}^{K}\left\|\mathbf{z}^k\right\|^2\right] \\
\leq& \frac{2\mathbb{E}\left[F(\mathbf{x}^1) - F(\mathbf{x}^{K+1})\right]}{\eta K} - \frac{1}{2K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{z}^k\right\|^2\right] + \frac{2\alpha T^2}{B_1}(\sigma^2 + L_f^2)^T + \frac{4L_F^2\eta^2}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{z}^k\right\|^2\right] \\
&+ 4TL_F^2\sum_{i=1}^{T-1}C_i\left(\frac{\eta^2}{\alpha^2 K}\sum_{k=1}^{K+1}\mathbb{E}\left[\left\|\mathbf{z}^k\right\|^2\right] + \frac{\sigma^2}{B_1}\right) \\
\leq& \frac{2\mathbb{E}\left[F(\mathbf{x}^1) - F(\mathbf{x}^{K+1})\right]}{\eta K} + \frac{2\alpha T^2}{B_1}(\sigma^2 + L_f^2)^T + \frac{\sigma^2}{B_1}4TL_F^2\sum_{i=1}^{T-1}C_i + \left(4TL_F^2\sum_{i=1}^{T-1}C_i\right)\frac{\eta^2}{\alpha^2 K}\mathbb{E}\left[\left\|\mathbf{z}^{K+1}\right\|^2\right] \\
\leq& \mathcal{O}\left(\frac{1}{\eta K} + \frac{1}{B_1} + \frac{\eta^2}{\alpha^2 K}\right) \leq \mathcal{O}\left(\frac{1}{K}\right).
\end{aligned}
$$

We complete the proof of the theorem by noticing $\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla F(\mathbf{x}^k)\right\|\right] \leq \sqrt{\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\mathbf{x}^k)\|^2\right]}$.

E PROOF OF THEOREM 3.8

By noting that $\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{\eta}{\|\mathbf{z}^k\|}\mathbf{z}^k$, we have:

$$\begin{aligned}
F\left(\mathbf{x}^{k+1}\right) \leq & F\left(\mathbf{x}^k\right) + \left\langle \nabla F(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \right\rangle + \frac{L_F}{2}\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2 \\
= & F(\mathbf{x}^k) - \eta \left\langle \nabla F(\mathbf{x}^k), \frac{\mathbf{z}^k}{\|\mathbf{z}^k\|} \right\rangle + \frac{\eta^2 L_F}{2} \\
= & F(\mathbf{x}^k) + \eta \left\langle \mathbf{z}^k - \nabla F(\mathbf{x}^k), \frac{\mathbf{z}^k}{\|\mathbf{z}^k\|} \right\rangle - \eta \left\langle \mathbf{z}^k, \frac{\mathbf{z}^k}{\|\mathbf{z}^k\|} \right\rangle + \frac{\eta^2 L_F}{2} \\
\leq & F(\mathbf{x}^k) + \eta \left\|\mathbf{z}^k - \nabla F(\mathbf{x}^k)\right\| \frac{\left\|\mathbf{z}^k\right\|}{\|\mathbf{z}^k\|} - \eta \frac{\left\|\mathbf{z}^k\right\|^2}{\|\mathbf{z}^k\|} + \frac{\eta^2 L_F}{2} \\
\leq & F(\mathbf{x}^k) + \eta \left\|\mathbf{z}^k - \nabla F(\mathbf{x}^k)\right\| - \eta \left\|\mathbf{z}^k\right\| + \frac{\eta^2 L_F}{2}.
\end{aligned}$$

This leads to the fact that

$$\left\|\nabla F(\mathbf{x}^k)\right\| \leq \left\|\mathbf{z}^k\right\| + \left\|\mathbf{z}^k - \nabla F(\mathbf{x}^k)\right\| \leq \frac{F(\mathbf{x}^k) - F(\mathbf{x}^{k+1})}{\eta} + 2\left\|\mathbf{z}^k - \nabla F(\mathbf{x}^k)\right\| + \frac{\eta L_F}{2}.$$

Summing up, by setting $\eta = 1/\sqrt{K}$, $B_0 = 1$, $B_1 = \lfloor K/4 \rfloor$, $\alpha = 1/2$ (actually α can be any positive constant within the range $(0, 1)$), we know that

$$\begin{aligned}
& \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\left\|\nabla F(\mathbf{x}^k)\right\|\right] \\
\leq & \frac{\mathbb{E}\left[F(\mathbf{x}^1) - F(\mathbf{x}^{K+1})\right]}{\eta K} + \mathbb{E}\left[\frac{2}{K}\sum_{k=1}^{K}\left\|\mathbf{z}^k - \nabla F(\mathbf{x}^k)\right\|\right] + \frac{\eta L_F}{2} \\
\leq & \frac{\mathbb{E}\left[F(\mathbf{x}^1) - F(\mathbf{x}^{K+1})\right]}{\eta K} + \frac{\eta L_F}{2} \\
& + 2\sqrt{\frac{2\alpha T^2}{B_1}(\sigma^2 + L_f^2)^T + \frac{4L_F^2}{\alpha^2 K}\sum_{k=1}^{K+1}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + 4TL_F^2\sum_{i=1}^{T-1} C_i\left(\frac{1}{\alpha^2 K}\sum_{k=1}^{K+1}\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2\right] + \frac{\sigma^2}{B_1}\right)} \\
\leq & \mathcal{O}\left(\frac{1}{\eta K} + \eta + \sqrt{\frac{\eta^2}{\alpha^2} + \frac{1}{B_1}}\right) \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right).
\end{aligned}$$

F PROOF OF LEMMA 4.1

Considering $\tilde{\mathbf{u}}_i^k = (1-\alpha)\left(\mathbf{u}_i^{k-1} + \langle \nabla f_i(\mathbf{u}_{i-1}^k;\xi_i^k), \mathbf{u}_{i-1}^k - \mathbf{u}_{i-1}^{k-1}\rangle\right) + \alpha f_i(\mathbf{u}_{i-1}^k;\xi_i^k)$, $\mathbf{u}_i^k = \mathrm{Clip}_C\left[\tilde{\mathbf{u}}_i^k\right]$ and $\left\|f_i(\mathbf{u}_{i-1}^k)\right\| \le C$, note that this operation can be viewed as a projection operation, and with the non-expansive property, we have:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{u}_i^k - f_i(\mathbf{u}_{i-1}^k)\right\|^2\right] \\
\le&\mathbb{E}\left[\left\|\tilde{\mathbf{u}}_i^k - f_i(\mathbf{u}_{i-1}^k)\right\|^2\right] \\
=&\mathbb{E}\left[\left\|(1-\alpha)\left(\mathbf{u}_i^{k-1} - f_i(\mathbf{u}_{i-1}^k) + \langle \nabla f_i(\mathbf{u}_{i-1}^k;\xi_i^k), \mathbf{u}_{i-1}^k - \mathbf{u}_{i-1}^{k-1}\rangle\right) + \alpha\left(f_i(\mathbf{u}_{i-1}^k;\xi_i^k) - f_i(\mathbf{u}_{i-1}^k)\right)\right\|^2\right] \\
=&\mathbb{E}\left[\left\|(1-\alpha)\left(\mathbf{u}_i^{k-1} - f_i(\mathbf{u}_{i-1}^k) + \langle \nabla f_i(\mathbf{u}_{i-1}^k), \mathbf{u}_{i-1}^k - \mathbf{u}_{i-1}^{k-1}\rangle\right)\right.\right. \\
&\left.\left.+(1-\alpha)\langle \nabla f_i(\mathbf{u}_{i-1}^k;\xi_i^k) - \nabla f_i(\mathbf{u}_{i-1}^k), \mathbf{u}_{i-1}^k - \mathbf{u}_{i-1}^{k-1}\rangle + \alpha\left(f_i(\mathbf{u}_{i-1}^k;\xi_i^k) - f_i(\mathbf{u}_{i-1}^k)\right)\right\|^2\right] \\
=&(1-\alpha)^2\mathbb{E}\left[\left\|\mathbf{u}_i^{k-1} - f_i(\mathbf{u}_{i-1}^k) + \langle \nabla f_i(\mathbf{u}_{i-1}^k), \mathbf{u}_{i-1}^k - \mathbf{u}_{i-1}^{k-1}\rangle\right\|^2\right] \\
&+\mathbb{E}\left[\left\|(1-\alpha)\langle \nabla f_i(\mathbf{u}_{i-1}^k;\xi_i^k) - \nabla f_i(\mathbf{u}_{i-1}^k), \mathbf{u}_{i-1}^k - \mathbf{u}_{i-1}^{k-1}\rangle + \alpha\left(f_i(\mathbf{u}_{i-1}^k;\xi_i^k) - f_i(\mathbf{u}_{i-1}^k)\right)\right\|^2\right] \\
\le&(1-\alpha)^2\mathbb{E}\left[\left\|\mathbf{u}_i^{k-1} - f_i(\mathbf{u}_{i-1}^{k-1})\right\|^2 + \left\|f_i(\mathbf{u}_{i-1}^{k-1}) - f_i(\mathbf{u}_{i-1}^k) + \langle \nabla f_i(\mathbf{u}_{i-1}^k), \mathbf{u}_{i-1}^k - \mathbf{u}_{i-1}^{k-1}\rangle\right\|^2\right] \\
&+(1-\alpha)^2\mathbb{E}\left[2\left\|\mathbf{u}_i^{k-1} - f_i(\mathbf{u}_{i-1}^{k-1})\right\|\left\|f_i(\mathbf{u}_{i-1}^{k-1}) - f_i(\mathbf{u}_{i-1}^k) + \langle \nabla f_i(\mathbf{u}_{i-1}^k), \mathbf{u}_{i-1}^k - \mathbf{u}_{i-1}^{k-1}\rangle\right\|\right] \\
&+2(1-\alpha)^2\mathbb{E}\left[\left\|\langle \nabla f_i(\mathbf{u}_{i-1}^k;\xi_i^k) - \nabla f_i(\mathbf{u}_{i-1}^k), \mathbf{u}_{i-1}^k - \mathbf{u}_{i-1}^{k-1}\rangle\right\|^2\right] + 2\alpha^2\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^k;\xi_i^k) - f_i(\mathbf{u}_{i-1}^k)\right\|^2\right] \\
\le&(1-\alpha)^2\mathbb{E}\left[\left\|\mathbf{u}_i^{k-1} - f_i(\mathbf{u}_{i-1}^{k-1})\right\|^2 + 2\left\|f_i(\mathbf{u}_{i-1}^{k-1}) - f_i(\mathbf{u}_{i-1}^k)\right\|^2 + 2\left\|\langle \nabla f_i(\mathbf{u}_{i-1}^k), \mathbf{u}_{i-1}^k - \mathbf{u}_{i-1}^{k-1}\rangle\right\|^2\right] \\
&+(1-\alpha)^2\mathbb{E}\left[L_J\left\|\mathbf{u}_i^{k-1} - f_i(\mathbf{u}_{i-1}^{k-1})\right\|\left\|\mathbf{u}_{i-1}^{k-1} - \mathbf{u}_{i-1}^k\right\|^2\right] \\
&+2\mathbb{E}\left[\left\|\nabla f_i(\mathbf{u}_{i-1}^k;\xi_i^k) - \nabla f_i(\mathbf{u}_{i-1}^k)\right\|^2\left\|\mathbf{u}_{i-1}^k - \mathbf{u}_{i-1}^{k-1}\right\|^2\right] + 2\alpha^2\sigma^2 \\
\le&(1-\alpha)^2\mathbb{E}\left[\left\|\mathbf{u}_i^{k-1} - f_i(\mathbf{u}_{i-1}^{k-1})\right\|^2 + 2L_f^2\left\|\mathbf{u}_{i-1}^{k-1} - \mathbf{u}_{i-1}^k\right\|^2 + 2L_f^2\left\|\mathbf{u}_{i-1}^k - \mathbf{u}_{i-1}^{k-1}\right\|^2\right] \\
&+(1-\alpha)^2\mathbb{E}\left[L_J\left(\left\|\mathbf{u}_i^{k-1}\right\| + \left\|f_i(\mathbf{u}_{i-1}^{k-1})\right\|\right)\left\|\mathbf{u}_{i-1}^{k-1} - \mathbf{u}_{i-1}^k\right\|^2\right] + 2\sigma^2\mathbb{E}\left[\left\|\mathbf{u}_{i-1}^k - \mathbf{u}_{i-1}^{k-1}\right\|^2\right] + 2\alpha^2\sigma^2 \\
\le&(1-\alpha)^2\mathbb{E}\left[\left\|\mathbf{u}_i^{k-1} - f_i(\mathbf{u}_{i-1}^{k-1})\right\|^2\right] + 4L_f^2\mathbb{E}\left[\left\|\mathbf{u}_{i-1}^{k-1} - \mathbf{u}_{i-1}^k\right\|^2\right] \\
&+2L_JC\mathbb{E}\left[\left\|\mathbf{u}_{i-1}^{k-1} - \mathbf{u}_{i-1}^k\right\|^2\right] + 2\sigma^2\mathbb{E}\left[\left\|\mathbf{u}_{i-1}^k - \mathbf{u}_{i-1}^{k-1}\right\|^2\right] + 2\alpha^2\sigma^2 \\
\le&(1-\alpha)\mathbb{E}\left[\left\|\mathbf{u}_i^{k-1} - f_i(\mathbf{u}_{i-1}^{k-1})\right\|^2\right] + (4L_f^2 + 2L_JC + 2\sigma^2)\mathbb{E}\left[\left\|\mathbf{u}_{i-1}^k - \mathbf{u}_{i-1}^{k-1}\right\|^2\right] + 2\alpha^2\sigma^2.
\end{aligned}
$$

Summing up and rearranging, we can obtain:

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^K\mathbb{E}\left[\left\|\mathbf{u}_i^k - f_i(\mathbf{u}_{i-1}^k)\right\|^2\right] \\
\le&\frac{\mathbb{E}\left[\left\|\mathbf{u}_i^1 - f_i(\mathbf{u}_{i-1}^1)\right\|^2\right]}{\alpha K} + \frac{(4L_f^2 + 2L_JC + 2\sigma^2)}{\alpha K}\sum_{k=1}^K\mathbb{E}\left[\left\|\mathbf{u}_{i-1}^{k+1} - \mathbf{u}_{i-1}^k\right\|^2\right] + 2\alpha\sigma^2 \\
\le&\frac{\sigma^2}{\alpha K} + \frac{(4L_f^2 + 2L_JC + 2\sigma^2)}{\alpha K}\sum_{k=1}^K\mathbb{E}\left[\left\|\mathbf{u}_{i-1}^{k+1} - \mathbf{u}_{i-1}^k\right\|^2\right] + 2\alpha\sigma^2 \\
\le&\frac{(4L_f^2 + 2L_JC + 2\sigma^2)}{\alpha K}\sum_{k=1}^K\mathbb{E}\left[\left\|\mathbf{u}_{i-1}^{k+1} - \mathbf{u}_{i-1}^k\right\|^2\right] + 3\alpha\sigma^2.
\end{aligned}
$$

where the last inequality is by setting $\alpha^2 K \ge 1$.

Next, we bound the difference of function estimators between adjacent steps:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{u}_i^{k+1}-\mathbf{u}_i^k\right\|^2\right] \leq \mathbb{E}\left[\left\|\tilde{\mathbf{u}}_i^{k+1}-\mathbf{u}_i^k\right\|^2\right] \\
\leq & 2\alpha^2\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^{k+1};\xi_i^{k+1})-\mathbf{u}_i^k\right\|^2\right]+2(1-\alpha)^2\mathbb{E}\left[\left\|\langle\nabla f_i(\mathbf{u}_{i-1}^{k+1};\xi_i^{k+1}),\mathbf{u}_{i-1}^{k+1}-\mathbf{u}_{i-1}^k\rangle\right\|^2\right] \\
\leq & 2\alpha^2\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^{k+1};\xi_i^{k+1})-f_i(\mathbf{u}_{i-1}^{k+1})+f_i(\mathbf{u}_{i-1}^{k+1})-\mathbf{u}_i^k\right\|^2\right] \\
&+2\mathbb{E}\left[\left\|\langle\nabla f_i(\mathbf{u}_{i-1}^{k+1};\xi_i^{k+1})-\nabla f_i(\mathbf{u}_{i-1}^{k+1}),\mathbf{u}_{i-1}^{k+1}-\mathbf{u}_{i-1}^k\rangle+\langle\nabla f_i(\mathbf{u}_{i-1}^{k+1}),\mathbf{u}_{i-1}^{k+1}-\mathbf{u}_{i-1}^k\rangle\right\|^2\right] \\
= & 2\alpha^2\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^{k+1};\xi_i^{k+1})-f_i(\mathbf{u}_{i-1}^{k+1})\right\|^2\right]+2\alpha^2\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^{k+1})-\mathbf{u}_i^k\right\|^2\right] \\
&+2\mathbb{E}\left[\left\|\langle\nabla f_i(\mathbf{u}_{i-1}^{k+1};\xi_i^{k+1})-\nabla f_i(\mathbf{u}_{i-1}^{k+1}),\mathbf{u}_{i-1}^{k+1}-\mathbf{u}_{i-1}^k\rangle\right\|^2\right]+2\mathbb{E}\left[\left\|\langle\nabla f_i(\mathbf{u}_{i-1}^{k+1}),\mathbf{u}_{i-1}^{k+1}-\mathbf{u}_{i-1}^k\rangle\right\|^2\right] \\
\leq & 2\alpha^2\sigma^2+2\alpha^2\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^{k+1})-f_i(\mathbf{u}_{i-1}^k)+f_i(\mathbf{u}_{i-1}^k)-\mathbf{u}_i^k\right\|^2\right] \\
&+2\mathbb{E}\left[\left\|\nabla f_i(\mathbf{u}_{i-1}^{k+1};\xi_i^{k+1})-\nabla f_i(\mathbf{u}_{i-1}^{k+1})\right\|^2\left\|\mathbf{u}_{i-1}^{k+1}-\mathbf{u}_{i-1}^k\right\|^2\right]+2\mathbb{E}\left[\left\|\nabla f_i(\mathbf{u}_{i-1}^{k+1})\right\|^2\left\|\mathbf{u}_{i-1}^{k+1}-\mathbf{u}_{i-1}^k\right\|^2\right] \\
\leq & 2\alpha^2\sigma^2+4\alpha^2\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^{k+1})-f_i(\mathbf{u}_{i-1}^k)\right\|^2\right]+4\alpha^2\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^k)-\mathbf{u}_i^k\right\|^2\right] \\
&+2\sigma^2\mathbb{E}\left[\left\|\mathbf{u}_{i-1}^{k+1}-\mathbf{u}_{i-1}^k\right\|^2\right]+2L_f^2\mathbb{E}\left[\left\|\mathbf{u}_{i-1}^{k+1}-\mathbf{u}_{i-1}^k\right\|^2\right] \\
\leq & 2\alpha^2\sigma^2+(2\sigma^2+6L_f^2)\mathbb{E}\left[\left\|\mathbf{u}_{i-1}^{k+1}-\mathbf{u}_{i-1}^k\right\|^2\right]+4\alpha^2\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^k)-\mathbf{u}_i^k\right\|^2\right].
\end{aligned}
$$

Summing up, we obtain that

$$
\frac{1}{K}\sum_{k=1}^K\mathbb{E}\left[\left\|\mathbf{u}_i^{k+1}-\mathbf{u}_i^k\right\|^2\right] \leq 2\alpha^2\sigma^2+\frac{6L_f^2+2\sigma^2}{K}\sum_{k=1}^K\mathbb{E}\left[\left\|\mathbf{u}_{i-1}^{k+1}-\mathbf{u}_{i-1}^k\right\|^2\right]+\frac{4\alpha^2}{K}\sum_{k=1}^K\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^k)-\mathbf{u}_i^k\right\|^2\right].
$$

G Proof of Lemma 4.2

Note that we have proved that

$$
\Gamma_i \leq \frac{(4L_f^2+2L_JC+2\sigma^2)}{\alpha}\Lambda_{i-1}+3\alpha\sigma^2; \quad \Lambda_i \leq 2\alpha^2\sigma^2+(6L_f^2+2\sigma^2)\Lambda_{i-1}+(4\alpha^2)\Gamma_i.
$$

Thus by setting $\alpha \leq 1/2$, we know that

$$
\begin{aligned}
\Gamma_i+\frac{\Lambda_i}{2\alpha} \leq & \frac{(4L_f^2+2L_JC+2\sigma^2)}{\alpha}\Lambda_{i-1}+3\alpha\sigma^2+\alpha\sigma^2+\frac{(3L_f^2+\sigma^2)\Lambda_{i-1}}{\alpha}+(2\alpha)\Gamma_i \\
\leq & \frac{(7L_f^2+2L_JC+3\sigma^2)}{\alpha}\Lambda_{i-1}+4\alpha\sigma^2+\Gamma_i.
\end{aligned}
$$

As a result, denoting $D=2(7L_f^2+2L_JC+3\sigma^2)$ and noting $\Lambda_0=\frac{1}{K}\sum_{k=1}^K\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\mathbf{x}^k\right\|^2\right]$, we obtain

$$
\begin{aligned}
\Lambda_i \leq & D\Lambda_{i-1}+8\alpha^2\sigma^2 \leq D^2\Lambda_{i-2}+8\alpha^2\sigma^2(1+D) \leq D^i\Lambda_0+8\alpha^2\sigma^2D^i \\
\leq & D^i\frac{1}{K}\sum_{k=1}^K\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\mathbf{x}^k\right\|^2\right]+8\alpha^2\sigma^2D^i.
\end{aligned}
$$

Then, we can obtain that

$$
\begin{aligned}
\Gamma_i \leq & \frac{D}{\alpha}\left[D^{i-1}\frac{1}{K}\sum_{k=1}^K\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\mathbf{x}^k\right\|^2\right]+8\alpha^2\sigma^2D^{i-1}\right]+3\alpha\sigma^2 \\
\leq & \frac{D^i}{\alpha}\frac{1}{K}\sum_{k=1}^K\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\mathbf{x}^k\right\|^2\right]+(8D^i+3)\alpha\sigma^2.
\end{aligned}
$$

H PROOF OF LEMMA 4.3

In the previous analysis, we have already proved that

$$\mathbb{E}\left[\left\|\mathbf{z}^{k+1}-\nabla F\left(\mathbf{x}^{k+1}\right)\right\|^2\right] \leq (1-\alpha)\mathbb{E}\left[\left\|\mathbf{z}^{k}-\nabla F(\mathbf{x}^{k})\right\|^2\right] + \frac{4L_F^2}{\alpha}\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\right\|^2\right] + 4\alpha T L_F^2 \sum_{i=1}^{T-1}\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^{k+1})-\mathbf{u}_i^{k+1}\right\|^2\right] + \alpha^2 T^2(\sigma^2+L_f^2)^T.$$

Summing up and rearranging,

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{z}^{k}-\nabla F(\mathbf{x}^{k})\right\|^2\right] \\
\leq&\frac{\mathbb{E}\left[\left\|\mathbf{z}^{1}-\nabla F(\mathbf{x}^{1})\right\|^2\right]}{\alpha K} + \frac{4L_F^2}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\right\|^2\right] + \frac{4TL_F^2}{K}\sum_{k=1}^{K}\sum_{i=1}^{T-1}\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^{k+1})-\mathbf{u}_i^{k+1}\right\|^2\right] + \alpha T^2(\sigma^2+L_f^2)^T \\
\leq&\alpha T^2(\sigma^2+L_f^2)^T + \frac{4L_F^2}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\right\|^2\right] + \frac{4TL_F^2}{K}\sum_{k=1}^{K+1}\sum_{i=1}^{T-1}\mathbb{E}\left[\left\|f_i(\mathbf{u}_{i-1}^{k})-\mathbf{u}_i^{k}\right\|^2\right] + \alpha T^2(\sigma^2+L_f^2)^T \\
\leq&2\alpha T^2(\sigma^2+L_f^2)^T + \frac{4L_F^2}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\right\|^2\right] + (4TL_F^2)\sum_{i=1}^{T-1}\left(\frac{D^i}{\alpha}\frac{1}{K}\sum_{k=1}^{K+1}\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\right\|^2\right] + (8D^i+3)\alpha\sigma^2\right) \\
\leq&2\alpha T^2(\sigma^2+L_f^2)^T + \frac{4L_F^2}{\alpha^2 K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\right\|^2\right] + \frac{4TL_F^2D^T}{\alpha}\frac{1}{K}\sum_{k=1}^{K+1}\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\right\|^2\right] + 4TL_F^2\sigma^2(8D^T+3T)\alpha \\
\leq&\left(2T^2(\sigma^2+L_f^2)^T + 4TL_F^2\sigma^2(8D^T+3T)\right)\alpha + 4L_F^2\left(1+TD^T\right)\frac{1}{\alpha^2 K}\sum_{k=1}^{K+1}\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\right\|^2\right] \\
\leq&D_0\alpha + 4L_F^2\left(1+TD^T\right)\frac{1}{\alpha^2 K}\sum_{k=1}^{K+1}\mathbb{E}\left[\left\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\right\|^2\right],
\end{aligned}$$

where $D_0 = 2T^2(\sigma^2+L_f^2)^T + 4TL_F^2\sigma^2(8D^T+3T)$.

I PROOF OF THEOREM 4.4

Due to the smoothness of function F, by setting $\eta \leq \frac{1}{2L_F}$, we have proved that $\mathbb{E}\left[\left\|\mathbf{z}^{k}\right\|^2\right] \leq 2(L_f^2+\sigma^2)^T$ and

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla F(\mathbf{x}^{k})\right\|^2\right] \\
\leq&\frac{2\mathbb{E}\left[F(\mathbf{x}^{1})-F(\mathbf{x}^{K+1})\right]}{\eta K} + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(\mathbf{x}^{k})-\mathbf{z}^{k}\right\|^2\right] - \frac{1}{2K}\mathbb{E}\left[\sum_{k=1}^{K}\left\|\mathbf{z}^{k}\right\|^2\right] \\
\leq&\frac{2\mathbb{E}\left[F(\mathbf{x}^{1})-F(\mathbf{x}^{K+1})\right]}{\eta K} - \frac{1}{2K}\mathbb{E}\left[\sum_{k=1}^{K}\left\|\mathbf{z}^{k}\right\|^2\right] + D_0\alpha + 4L_F^2\left(1+TD^T\right)\frac{\eta^2}{\alpha^2 K}\sum_{k=1}^{K+1}\mathbb{E}\left[\left\|\mathbf{z}^{k}\right\|^2\right] \\
\leq&\mathcal{O}\left(\frac{1}{\eta K}+\alpha+\frac{\eta^2}{\alpha^2 K}\right) \leq \mathcal{O}\left(\frac{1}{\sqrt{K}}\right),
\end{aligned}$$

where we set $\alpha = \mathcal{O}(K^{-1/2})$ and learning rate $\eta \leq \frac{\alpha}{2L_F\sqrt{2(1+TD^T)}} = \mathcal{O}(K^{-1/2})$. We complete the proof of the theorem by noticing $\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla F(\mathbf{x}^{k})\right\|\right] \leq \sqrt{\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla F(\mathbf{x}^{k})\right\|^2\right]}$.

J Dependence Issues in Previous Literature

The main problem in the analysis of Balasubramanian et al. (2022) lies in the variable dependence in the analysis. For example, let us examine their proof for Proposition 3.4 (refer to Proposition 3.1 if using the arXiv version). In the proof, they use

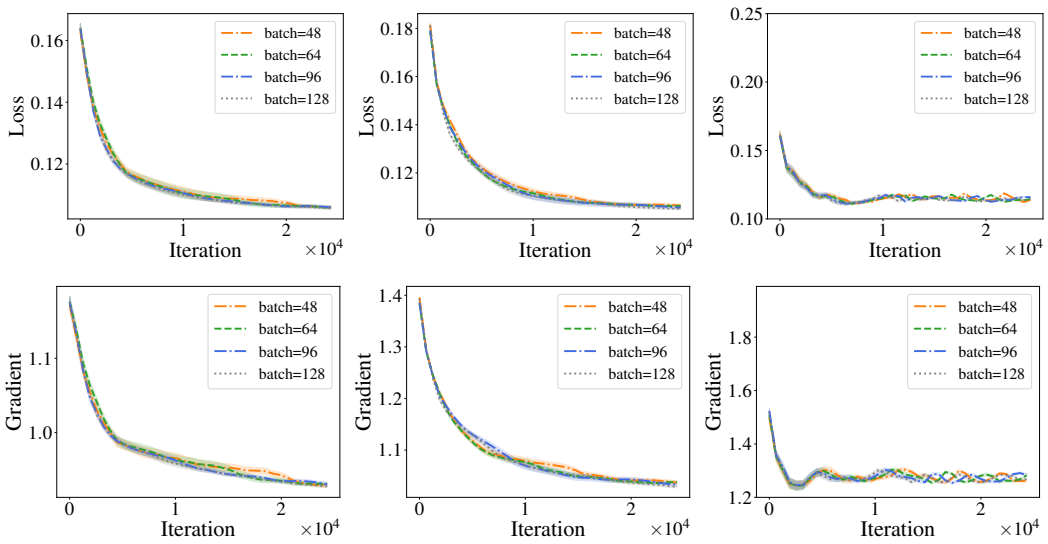
$$\begin{aligned}&\mathbb{E}\left[\left\|f_i\left(w_{i-1}^{k+1}\right)-w_i^{k+1}\right\|\left\|w_i^{k+1}-w_i^k\right\|^2|\mathcal{F}_k\right]\\&\leq\left(\mathbb{E}\left[\left\|f_i\left(w_{i-1}^{k+1}\right)-w_i^{k+1}\right\|^2|\mathcal{F}_k\right]\right)^{\frac{1}{2}}\mathbb{E}\left[\left\|w_i^{k+1}-w_i^k\right\|^2|\mathcal{F}_k\right].\end{aligned}$$

This inequality is not true since we cannot ensure $\mathbb{E}[AB] \leq \mathbb{E}[A]\cdot\mathbb{E}[B]$ when A and B are not independent (e.g., when $A = B = X$ and $\mathbb{E}[X] = 0$, it is obvious that we can not ensure $\mathbb{E}[X^2] \leq \mathbb{E}[X]\mathbb{E}[X] = 0$). Note that w_i^{k+1} appears in both terms. Thus, we cannot ensure

$$\begin{aligned}&\mathbb{E}\left[\left\|f_i\left(w_{i-1}^{k+1}\right)-w_i^{k+1}\right\|\left\|w_i^{k+1}-w_i^k\right\|^2|\mathcal{F}_k\right]\\&\leq\mathbb{E}\left[\left\|f_i\left(w_{i-1}^{k+1}\right)-w_i^{k+1}|\mathcal{F}_k\right\|\right]\mathbb{E}\left[\left\|w_i^{k+1}-w_i^k\right\|^2|\mathcal{F}_k\right]\\&\leq\left(\mathbb{E}\left[\left\|f_i\left(w_{i-1}^{k+1}\right)-w_i^{k+1}\right\|^2|\mathcal{F}_k\right]\right)^{\frac{1}{2}}\mathbb{E}\left[\left\|w_i^{k+1}-w_i^k\right\|^2|\mathcal{F}_k\right].\end{aligned}$$

As a result, the analysis is invalid without this inequality.

K Additional Experimental Results



(a) Industry-10 (b) Industry-12 (c) Industry-17

Figure 3: Results for Risk-Averse Portfolio Optimization with different batch sizes.

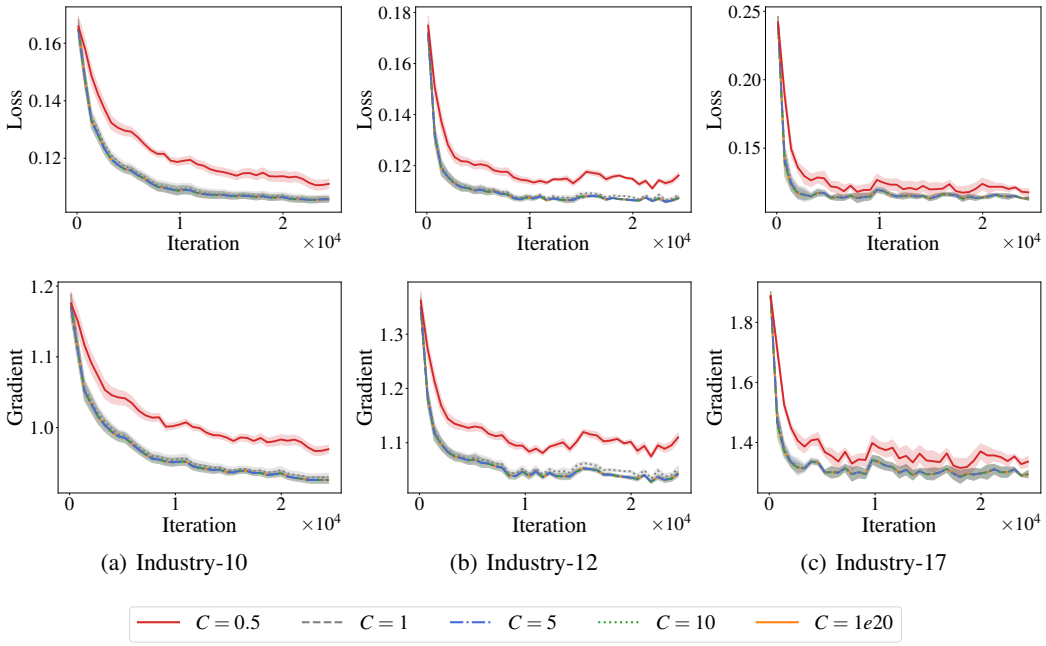


Figure 4: Sensitivity to the clipping threshold C on the portfolio task.

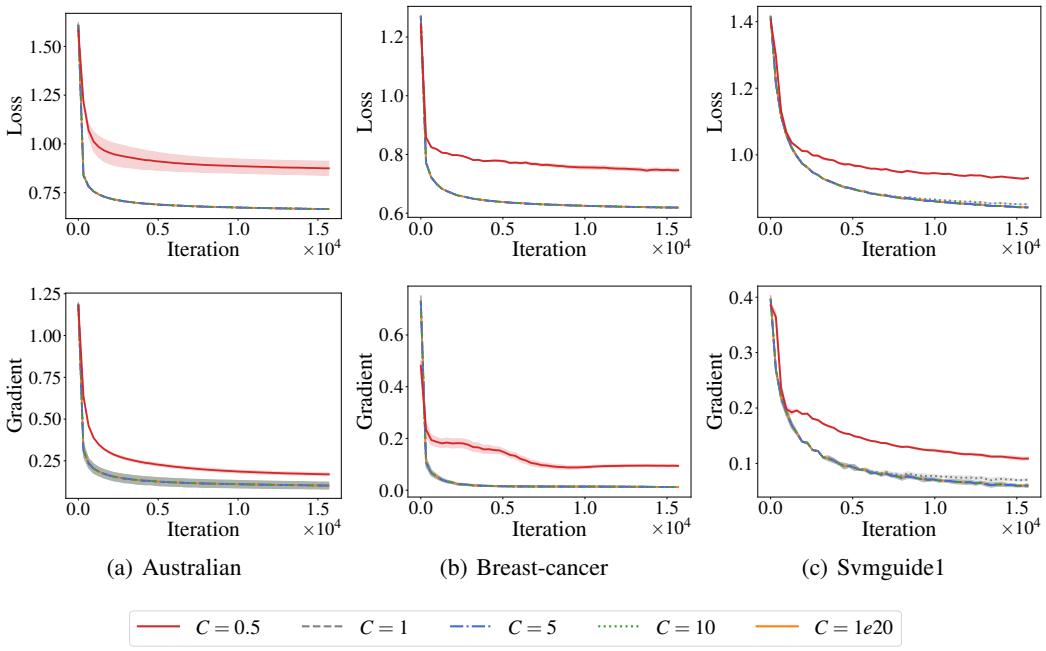


Figure 5: Sensitivity to the clipping threshold C on hierarchical TERM.