FEDNEST: Federated Bilevel Optimization

Davoud Ataee Tarzanagh [1] **Mingchen Li** [2] **Christos Thrampoulidis** [3] **Samet Oymak** [2]

Abstract

Standard federated optimization methods are being successfully applied to solve stochastic problems with a *single-level* structure. However, many contemporary ML problems – including adversarial robustness, hyperparameter tuning, actor-critic – fall under nested bilevel programming that subsumes compositional and min-max optimization. In this work, we propose FEDNEST: A federated alternating stochastic gradient method to address general nested problems. We establish provable convergence rates for FEDNEST in the presence of heterogeneous data and introduce variations for specific instances. FEDNEST introduces multiple innovations including federated hypergradient computation and variance reduction to address inner-level heterogeneity. We complement our theory with experiments on hyperparameter & hyper-representation learning that demonstrate the benefits of our method in practice.

1. Introduction

In the Federated learning (FL) paradigm, multiple clients cooperate to learn a model under the orchestration of a central server (McMahan et al., 2017) without directly exchanging local client data with the server or other clients. The locality of data distinguishes FL from traditional distributed optimization and also motivates new methodologies to address heterogeneous data across clients. Additionally, cross-device FL across many edge devices presents additional challenges since only a small fraction of clients participate in each round, and clients cannot maintain *state* across rounds Kairouz et al. (2019).

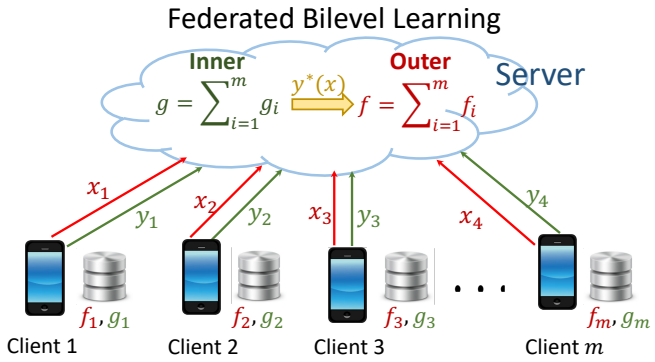
Traditional distributed SGD methods are often unsuitable in FL and incur high communication costs. To overcome this issue, popular FL methods, such as FEDAVG (McMahan et al., 2017) ,use local client updates, i.e. clients update their models multiple times before communicating with the server (aka, local SGD). Although FEDAVG has seen great success, recent works have exposed convergence issues in certain settings (Karimireddy et al., 2020; Hsu et al., 2019). This is due to a variety of factors, including *client drift* (Karimireddy et al., 2020), where local models move away from globally optimal models due to objective and/or systems heterogeneity.

Existing federated optimization methods, such as FEDAVG, are widely applied to stochastic problems with *single-level* structure. Instead, many machine learning tasks – such as adversarial learning, meta learning (Franceschi et al., 2018; Bertinetto et al., 2018), hyperparameter optimization (Franceschi et al., 2018; Feurer & Hutter, 2019), reinforcement/imitation learning (Arora et al., 2020; Hong et al., 2020), and neural architecture search (Liu et al., 2018) – admit *nested* formulations that go beyond the standard single-level structure. In an attempt to address such nested problems, bilevel optimization, as well as its special cases min-max and composite optimization, have received significant attention in the recent literature (Ghadimi & Wang, 2018; Liu et al., 2021; Chen et al., 2021a; Hong et al., 2020; Lin et al., 2020; Rafique et al., 2021; Ji et al., 2020a; Lorraine et al., 2020); albeit in non-federated settings. On the other hand, federated versions have been elusive perhaps due to the additional challenges surrounding heterogeneity, communication, and inverse Hessian estimation.

Contributions: This paper addresses these challenges and develops **FEDNEST**: A federated machinery for the stochastic bilevel problems with provable convergence and lightweight communication. FEDNEST is composed of **FEDINN**: a federated stochastic variance reduction algorithm (SVRG) for solving the inner problem, and **FEDOUT**: a communication-efficient federated hypergradient algorithm for solving the outer problem. Importantly, we allow both inner and outer objectives to be finite sums over heterogeneous local client functions. FEDNEST runs a variant of federated SVRG on inner & outer variables in an alternating fashion as outlined in Algo. 1. We make multiple algorithmic and theoretical contributions summarized below.

- **The variance reduction** of FEDINN enables robustness in the sense that local models converge to the globally optimal inner model *despite client drift/heterogeneity*

[1]Email: tarzanaq@umich.edu, University of Michigan [2]Emails: {mli176@,oymak@ece.}ucr.edu, University of California, Riverside [3]Email: cthrampo@ece.ubc.ca, University of British Columbia.



Communication-efficiency:
- ☑ FEDIHGP avoids explicit Hessian
- ☑ LFEDNEST for local hypergradients

Client heterogeneity:
- ☑ FEDINN avoids client drift

Finite sample bilevel theory:
- ☑ Stochastic inner & outer analysis

Special nested problems:
- ☑ Min-max optimization
- ☑ Compositional optimization

Figure 1: Depiction of federated bilevel optimization and high-level summary FEDNEST (Algo. 1). At outer loop, FEDIHGP uses multiple rounds of Hessian-vector products to facilitate hypergradient computation while only communicating vectors. At inner loop, FEDINN uses SVRG to avoid client drift and find the unique global minima under inner strong-convexity (Konečnỳ et al., 2018; Mitra et al., 2021). Both are crucial for establishing provable convergence of FEDNEST.

(unlike FEDAVG). While FEDINN is similar to federated SVRG (Konečnỳ et al., 2018) and its improved variant FEDLIN (Mitra et al., 2021), we make two key contributions: (i) We leverage the global convergence of FEDINN to ensure accurate hypergradient computation which is crucial for our *bilevel* proof. (ii) We establish new convergence guarantees for *single-level* stochastic non-convex federated SVRG, which are then integrated within our FEDOUT.

- **Communication efficiency:** Within FEDOUT, we develop an efficient federated method for hypergradient estimation that bypass Hessian computation. Our approach is based on approximating the global Inverse Hessian-Gradient-Product (IHGP) via computation of Hessian-vector products over few communication rounds.
- **LFEDNEST:** To further improve communication efficiency, we additionally propose a Light-FEDNEST algorithm, which computes hypergradients locally and *only needs a single communication round* for the outer update.
- **Unified federated nested theory:** We specialize our bilevel results to *min-max & compositional* optimization. For these, FEDNEST significantly simplifies (no need for IHGP) and leads to faster convergence. Importantly, our results are *on par with the state-of-the-art non-federated guarantees* for nested optimization literature *without additional assumptions* (Tables 1 and 2).
- **Numerical experiments** on hyper-parameter tuning and hyper-representation (bilevel problems) demonstrate the benefits of FEDNEST, efficiency of LFEDNEST, and shed light on tradeoffs surrounding communication, computation, heterogeneity.

2. Federated Bilevel Optimization & FEDNEST

We will first provide the background on stochastic bilevel problems and then introduce our general federated method.

Stochastic Bilevel Optimization

		Non-Federated		
	FEDNEST	**ALSET**	**BSA**	**TTSA**
batch size	$\mathcal{O}(1)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$
samples in ξ	$\mathcal{O}(\kappa_g^5\epsilon^{-2})$	$\mathcal{O}(\kappa_g^5\epsilon^{-2})$	$\mathcal{O}(\kappa_g^6\epsilon^{-2})$	$\mathcal{O}(\kappa_g^p\epsilon^{-2.5})$
samples in ζ	$\mathcal{O}(\kappa_g^9\epsilon^{-2})$	$\mathcal{O}(\kappa_g^9\epsilon^{-2})$	$\mathcal{O}(\kappa_g^9\epsilon^{-3})$	$\mathcal{O}(\kappa_g^p\epsilon^{-2.5})$

Stochastic Min-Max Optimization

		Non-Federated		
	FEDNEST	**ALSET**	**SGDA**	**SMD**
batch size	$\mathcal{O}(1)$	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-1})$	N.A.
samples	$\mathcal{O}(\kappa_f^3\epsilon^{-2})$	$\mathcal{O}(\kappa_f^3\epsilon^{-2})$	$\mathcal{O}(\kappa_f^3\epsilon^{-2})$	$\mathcal{O}(\kappa_f^3\epsilon^{-2})$

Table 1: Sample complexity of FEDNEST and comparable non-federated methods to find an ϵ-stationary point of f. Here, $\kappa_g := \ell_{g,1}/\mu_g$ and $\kappa_f := \ell_{f,1}/\mu_f$. The notation κ_g^p denotes a polynomial function of κ_g. ALSET (Chen et al., 2021a), BSA (Ghadimi & Wang, 2018), TTSA (Hong et al., 2020), SGDA (Lin et al., 2020), SMD (Rafique et al., 2021).

Notation. For a differentiable function $h(\boldsymbol{x}, \boldsymbol{y}) : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ in which $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}) : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$, we denote $\nabla h \in \mathbb{R}^{d_1}$ the gradient of h as a function of $\boldsymbol{x}$ and $\nabla_{\boldsymbol{x}} h$, $\nabla_{\boldsymbol{y}} h$ the partial derivatives of h with respect to $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. We let $\nabla^2_{\boldsymbol{xy}} h$ and $\nabla^2_{\boldsymbol{y}} h$ denote the Jacobian and Hessian of h, respectively. We consider FL optimization over m clients and we denote $\mathcal{S} = \{1, \ldots, m\}$.

2.1. Preliminaries

In **federated bilevel learning**, we consider the following nested optimization problem as depicted in Figure 1:

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^{d_1}} \quad & f(\boldsymbol{x}) = \tfrac{1}{m} \textstyle\sum_{i=1}^m f_i\left(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})\right) \\ \text{subj. to} \quad & \boldsymbol{y}^*(\boldsymbol{x}) \in \operatorname*{argmin}_{\boldsymbol{y} \in \mathbb{R}^{d_2}} \tfrac{1}{m} \textstyle\sum_{i=1}^m g_i\left(\boldsymbol{x}, \boldsymbol{y}\right). \end{aligned} \tag{1a}$$

Recall that m is the number of clients. Here, to model objective heterogeneity, each client i is allowed to have its own individual outer & inner functions (f_i, g_i). Moreover, we consider a general stochastic oracle model, access to

Algorithm 1 FEDNEST

Inputs: $K, T \in \mathbb{N}$; $(\boldsymbol{x}^0, \boldsymbol{y}^0) \in \mathbb{R}^{d_1+d_2}$; FEDINN, FEDOUT with stepsizes $\{(\alpha^k, \beta^k)\}_{k=0}^{K-1}$
for $k = 0, \cdots, K-1$ **do**
  $\boldsymbol{y}^{k,0} = \boldsymbol{y}^k$
  for $t = 0, \cdots, T-1$ **do**
    $\boldsymbol{y}^{k,t+1} = \text{FEDINN}\left(\boldsymbol{x}^k, \boldsymbol{y}^{k,t}, \beta^k\right)$
  end for
  $\boldsymbol{y}^{k+1} = \boldsymbol{y}^{k,T}$
  $\boldsymbol{x}^{k+1} = \text{FEDOUT}\left(\boldsymbol{x}^k, \boldsymbol{y}^{k+1}, \alpha^k\right)$
end for

local functions (f_i, g_i) is via stochastic sampling as follows:

$$f_i(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})) := \mathbb{E}_{\xi\sim\mathcal{C}_i}[f_i(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \xi)],$$
$$g_i(\boldsymbol{x}, \boldsymbol{y}) := \mathbb{E}_{\zeta\sim\mathcal{D}_i}[g_i(\boldsymbol{x}, \boldsymbol{y}; \zeta)], \tag{1b}$$

where, $(\xi, \zeta) \sim (\mathcal{C}_i, \mathcal{D}_i)$ are outer/inner sampling distributions for the i^{th} client. We emphasize that for $i \neq j$, the tuples $(f_i, g_i, \mathcal{C}_i, \mathcal{D}_i)$ and $(f_j, g_j, \mathcal{C}_j, \mathcal{D}_j)$ can be different.

Example 1 (Hyperparameter tuning)**.** *Each client has local validation and training datasets associated with objectives $(f_i, g_i)_{i=1}^m$ corresponding to validation and training losses, respectively. The goal is inding hyper-parameters $\boldsymbol{x}$ that lead to learning model parameters $\boldsymbol{y}$ that minimize the (global) validation loss.*

The *stochastic bilevel* problem (1) subsumes two popular problem classes with the nested structure: *Stochastic Min-max & Stochastic Compositional*. Therefore, results on the general nested problem (1) also imply the results in these special cases. Below, we briefly describe them.

Min-max optimization. If $g_i(\boldsymbol{x}, \boldsymbol{y}; \zeta) := -f_i(\boldsymbol{x}, \boldsymbol{y}; \xi)$ for all $i \in [m]$, the stochastic bilevel problem (1) reduces to the stochastic min-max problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^{d_1}} f(\boldsymbol{x}) := \frac{1}{m} \max_{\boldsymbol{y}\in\mathbb{R}^{d_2}} \sum_{i=1}^{m} \mathbb{E}[f_i(\boldsymbol{x}, \boldsymbol{y}; \xi)]. \tag{2}$$

Motivated by applications in fair beamforming, training generative-adversarial networks (GANs) and robust machine learning, significant efforts have been made for solving (2) including (Daskalakis & Panageas, 2018; Gidel et al., 2018; Mokhtari et al., 2020; Thekumparampil et al., 2019).

Example 2 (GANs)**.** *We train a generative model $g_{\boldsymbol{x}}(\cdot)$ and an adversarial model $a_{\boldsymbol{y}}(\cdot)$ using client datasets $\mathcal{S}_i$. The local functions may for example take the form $f_i(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_{s\sim\mathcal{S}_i}[\log a_{\boldsymbol{y}}(s)] + \mathbb{E}_{z\sim\mathcal{D}_{noise}}[\log(1 - a_{\boldsymbol{y}}(g_{\boldsymbol{x}}(z)))]$.*

Compositional optimization. Suppose $f_i(\boldsymbol{x}, \boldsymbol{y}; \xi) := f_i(\boldsymbol{y}; \xi)$ and g_i is quadratic in $\boldsymbol{y}$ given as $g_i(\boldsymbol{x}, \boldsymbol{y}; \zeta) := \|\boldsymbol{y} - \boldsymbol{h}_i(\boldsymbol{x}; \zeta)\|^2$. Then, (1) reduces to

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^{d_1}} \quad & f(\boldsymbol{x}) = \tfrac{1}{m}\textstyle\sum_{i=1}^m f_i\left(\boldsymbol{y}^*(\boldsymbol{x})\right) \\ \text{subj. to} \quad & \boldsymbol{y}^*(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{y}\in\mathbb{R}^{d_2}} \tfrac{1}{m}\textstyle\sum_{i=1}^m g_i\left(\boldsymbol{x}, \boldsymbol{y}\right). \end{aligned} \tag{3}$$

with $f_i(\boldsymbol{y}^*(\boldsymbol{x})) := \mathbb{E}_{\xi\sim\mathcal{C}_i}[f_i(\boldsymbol{y}^*(\boldsymbol{x}); \xi)]$ and $g_i(\boldsymbol{x}, \boldsymbol{y}) := g_i(\boldsymbol{x}, \boldsymbol{y}; \zeta)$. Stochastic compositional optimization problems in the form of (3) occur for example in model agnostic meta-learning and policy evaluation in reinforcement learning (Finn et al., 2017; Ji et al., 2020b; Dai et al., 2017; Wang et al., 2017; Chen et al., 2021a).

Assumptions. Let $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{d_1+d_2}$. Throughout, we make the following assumptions on inner/outer objectives.

Assumption A (Well-behaved objectives)**.** *For all $i \in [m]$:*

(A1) *$f_i(\boldsymbol{z}), \nabla f_i(\boldsymbol{z}), \nabla g_i(\boldsymbol{z}), \nabla^2 g_i(\boldsymbol{z})$ are $\ell_{f,0}$,$\ell_{f,1}$,$\ell_{g,1}$, $\ell_{g,2}$-Lipschitz continuous, respectively; and*

(A2) *$g_i(\boldsymbol{x}, \boldsymbol{y})$ is μ_g-strongly convex in $\boldsymbol{y}$ for all $\boldsymbol{x} \in \mathbb{R}^{d_1}$.*

Throughout, we use $\kappa_g = \ell_{g,1}/\mu_g$ to denote the condition number of the inner function g.

Assumption B (Stochastic samples)**.** *For all $i \in [m]$:*

(B1) *$\nabla f_i(\boldsymbol{z}; \xi)$, $\nabla g_i(\boldsymbol{z}; \zeta)$, $\nabla^2 g_i(\boldsymbol{z}; \zeta)$ are unbiased estimators of $\nabla f_i(\boldsymbol{z})$, $\nabla g_i(\boldsymbol{z})$, $\nabla^2 g_i(\boldsymbol{z})$, respectively; and*

(B2) *Their variances are bounded, i.e., $\mathbb{E}_\xi[\|\nabla f_i(\boldsymbol{z}; \xi) - \nabla f_i(\boldsymbol{z})\|^2] \leq \sigma_f^2$, $\mathbb{E}_\zeta[\|\nabla^2 g_i(\boldsymbol{z}; \zeta) - \nabla^2 g_i(\boldsymbol{z})\|^2] \leq \sigma_{g,1}^2$, and $\mathbb{E}_\zeta[\|\nabla^2 g_i(\boldsymbol{z}; \zeta) - \nabla^2 g_i(\boldsymbol{z})\|^2] \leq \sigma_{g,2}^2$ for some $\sigma_f^2, \sigma_{g,1}^2$, and $\sigma_{g,2}^2$.*

These assumptions are common in the bilevel optimization literature (Ghadimi & Wang, 2018; Hong et al., 2020; Chen et al., 2021a; Ji et al., 2021). Assumption A requires that the inner and outer functions are well-behaved. Specifically, strong-convexity of the inner objective is a recurring assumption in bilevel optimization theory implying a unique solution to the inner minimization in (1); see, e.g., (Ghadimi & Wang, 2018; Lorraine et al., 2020; Liu et al., 2021). Note that Assumptions A&B yield the following upper bound on the second moments, for all $i \in \{1, \ldots, m\}$:

$$\mathbb{E}_\xi[\|\nabla f_i(\boldsymbol{x}, \boldsymbol{y}; \xi)\|^2] \leq \ell_{f,0}^2 + \sigma_f^2 =: C_f^2, \tag{4a}$$
$$\mathbb{E}_\zeta[\|\nabla^2 g_i(\boldsymbol{x}, \boldsymbol{y}; \zeta)\|^2] \leq \ell_{g,1}^2 + \sigma_{g,2}^2 =: C_g^2. \tag{4b}$$

2.2. Proposed Algorithm: FEDNEST

In this section, we develop FEDNEST, which is formally presented in Algorithm 1. The algorithm operates in two nested loops. The outer loop operates in rounds $k \in \{1, \ldots, K\}$. Within each round, an inner loop operating for T iterations

is executed. Given estimates $\boldsymbol{x}^k$ and $\boldsymbol{y}^k$, each iteration $t \in \{1, \dots, T\}$ of the inner loop produces a new *global* model $\boldsymbol{y}^{k,t+1}$ of the inner optimization variable $\boldsymbol{y}^*(\boldsymbol{x}^k)$ as the output of an optimizer FEDINN. The final estimate $\boldsymbol{y}^{k+1} = \boldsymbol{y}^{k,T}$ of the inner variable is then used by an optimizer FEDOUT to update the outer *global* model $\boldsymbol{x}^{k+1}$.

The subroutines FEDINN and FEDOUT are gradient-based optimizers. Each subroutine involves a certain number of *local* training steps indexed by $\ell \in \{1, \dots, \tau_i\}$ that are performed at the i^{th} client. The local steps of FEDINN iterate over *local* models $\boldsymbol{y}_{i,\ell}$ of the inner variable. Accordingly, FEDOUT iterates over *local* models $\boldsymbol{x}_{i,\ell}$ of the global variable. A critical component of FEDOUT is a communication-efficient federated hypergradient optimization routine, which we call FEDIHGP. The implementation of FEDINN, FEDOUT and FEDIHGP is critical to circumvent the algorithmic challenges of FL bilevel optimization. In the remaining of this section, we detail the challenges and motivate our proposed implementations. Later, in Section 3, we provide a formal convergence analysis of FEDNEST.

2.3. Key Challenge: Federated Hypergradient Estimation

FEDOUT is a *gradient-based* optimizer for the outer minimization in (1); thus each iteration involves computing $\nabla f(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^m \nabla f_i(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}))$. Unlike single-level FL, the fact that the outer objective f depends explicitly on the inner minimizer $\boldsymbol{y}^*(\boldsymbol{x})$ introduces a new challenge. A good starting point to understand the challenge is the following evaluation of $\nabla f(\boldsymbol{x})$ in terms of partial derivatives. The result is well-known from properties of implicit functions.

Lemma 2.1. *Under Assumption A, for all $i \in [m]$:*

$$\nabla f_i(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})) = \nabla^{\text{D}} f_i\left(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})\right) + \nabla^{\text{I}} f_i\left(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})\right),$$

where the direct *and* indirect *gradient components are:*

$$\nabla^{\text{D}} f_i(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})) := \nabla_{\boldsymbol{x}} f_i\left(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})\right), \tag{5a}$$

$$\nabla^{\text{I}} f_i(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})) := -\nabla^2_{\boldsymbol{xy}} g(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})) \cdot [\nabla^2_{\boldsymbol{y}} g(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}))]^{-1} \nabla_{\boldsymbol{y}} f_i\left(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})\right). \tag{5b}$$

We now use the above formula to describe the two core challenges of bilevel FL optimization.

First, evaluation of any of the terms in (5) requires access to the minimizer $\boldsymbol{y}^*(\boldsymbol{x})$ of the inner problem. On the other hand, one may at best hope for a good *approximation* to $\boldsymbol{y}^*(\boldsymbol{x})$ produced by the inner optimization subroutine. Of course, this challenge is inherent in any bilevel optimization setting, but is exacerbated in the FL setting because of *client drift*. Specifically, when clients optimize their individual (possibly different) local inner objectives, the global estimate of the inner variable produced by SGD-type methods may drift far from (a good approximation to) $\boldsymbol{y}^*(\boldsymbol{x})$. We explain in Section 2.5 how FEDINN solves that issue.

The second challenge comes from the stochastic nature of the problem. Observe that the indirect component in (5b) is nonlinear in the Hessian $\nabla^2_{\boldsymbol{y}} g(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}))$, complicating an unbiased stochastic approximation of $\nabla f_i(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}))$. As we expose here, solutions to this complication developed in the non-federated bilevel optimization literature, are *not* directly applicable in the FL setting. Indeed, existing stochastic bilevel algorithms, e.g. (Ghadimi & Wang, 2018; Hong et al., 2020), define $\bar{\nabla} f(\boldsymbol{x}, \boldsymbol{y}) := \bar{\nabla}^{\text{D}} f(\boldsymbol{x}, \boldsymbol{y}) + \bar{\nabla}^{\text{I}} f(\boldsymbol{x}, \boldsymbol{y})$ as a surrogate of $\nabla f(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}))$ by replacing $\boldsymbol{y}^*(\boldsymbol{x})$ in definition (5) with an approximation $\boldsymbol{y}$ and using the following stochastic approximations:

$$\bar{\nabla}^{\text{D}} f(\boldsymbol{x}, \boldsymbol{y}) \approx \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}; \dot{\xi}), \tag{6a}$$

$$\bar{\nabla}^{\text{I}} f(\boldsymbol{x}, \boldsymbol{y}) \approx -\nabla^2_{\boldsymbol{xy}} g(\boldsymbol{x}, \boldsymbol{y}; \zeta_{N'+1}) \Big[\frac{N}{\ell_{g,1}} \prod_{n=1}^{N'} \big(\boldsymbol{I} - \frac{1}{\ell_{g,1}} \nabla^2_{\boldsymbol{y}} g(\boldsymbol{x}, \boldsymbol{y}; \zeta_n)\big)\Big] \nabla_y f(\boldsymbol{x}, \boldsymbol{y}; \dot{\xi}). \tag{6b}$$

Here, N' is drawn from $\{0, \dots, N-1\}$ uniformly at random (UAR) and $\{\dot{\xi}, \zeta_1, \dots, \zeta_{N'}\}$ are i.i.d. samples. Ghadimi & Wang (2018); Hong et al. (2020) have shown that using (6a), the inverse Hessian estimation bias exponentially decreases with the number of samples N. One might hope to directly leverage the above approach in a local computation fashion by replacing the global outer function f with the individual function f_i. However, note from (6b) that the proposed stochastic approximation of the indirect gradient involves in a nonlinear way the *global* Hessian, which is *not* available at the client [1]. Communication efficiency is one of the core objectives of FL making the idea of communicating Hessians between clients and server prohibitive. *Is it then possible, in a FL setting, to obtain an accurate stochastic estimate of the (local) indirect gradient while retaining communication efficiency?* In Section 2.4, we show how FEDOUT and its subroutine FEDIHGP, a gradient-based (thus, communication efficient) Federated hypergradient estimator, answer this question affirmatively.

2.4. FEDOUT

This section presents the outer optimizer FEDOUT, formally described in Algorithm 2. As a subroutine of FEDNEST (see Line 9, Alg. 1), at each round $k = 0, \dots, K-1$, FEDOUT takes the most recent global outer model $\boldsymbol{x}^k$ together the updated (by FEDINN) global inner model $\boldsymbol{y}^{k+1}$ and produces an update $\boldsymbol{x}^{k+1}$. To lighten notation, for a round k, denote

[1] We note that the approximation in (6a) is not the only construction, and bilevel optimization can accommodate other forms of gradient surrogates (Ji et al., 2021). Yet, all these approximations require access (in a nonlinear fashion) to the *global* Hessian; thus, they suffer from the same challenge in FL setting.

Algorithm 2 $x^+ =$ **FEDOUT** (x, y^+, α) for stochastic bilevel and min-max problems

1: $\mathbb{F}_i(\cdot) \leftarrow \nabla_{\boldsymbol{x}} f_i(\cdot, \boldsymbol{y}^+; \cdot)$
2: $\boldsymbol{x}_{i,0} = \boldsymbol{x}$ and $\alpha_i \in (0, \alpha]$
3: Choose $N \geq 1$ and set $\boldsymbol{p}_N =$ **FEDIHGP** $(\boldsymbol{x}, \boldsymbol{y}^+, N)$
4: **for** $i \in \mathcal{S}$ **in parallel do**
5: $\quad \boldsymbol{h}_i = \mathbb{F}_i(\boldsymbol{x}; \xi_i) - \nabla^2_{\boldsymbol{xy}} g_i(\boldsymbol{x}, \boldsymbol{y}^+; \zeta_i) \boldsymbol{p}_N$
6: $\quad \boldsymbol{h}_i = \mathbb{F}_i(\boldsymbol{x}; \xi_i)$
7: **end for**
8: $\boldsymbol{h} = |\mathcal{S}|^{-1} \sum_{i \in \mathcal{S}} \boldsymbol{h}_i$
9: **for** $i \in \mathcal{S}$ **in parallel do**
10: $\quad$ **for** $\ell = 0, \ldots, \tau_i - 1$ **do**
11: $\quad\quad \boldsymbol{h}_{i,\ell} = \mathbb{F}_i(\boldsymbol{x}_{i,\ell}; \xi_{i,\ell}) - \mathbb{F}_i(\boldsymbol{x}; \xi_{i,\ell}) + \boldsymbol{h}$
12: $\quad\quad \boldsymbol{x}_{i,\ell+1} = \boldsymbol{x}_{i,\ell} - \alpha_i \boldsymbol{h}_{i,\ell}$
13: $\quad$ **end for**
14: **end for**
15: $\boldsymbol{x}^+ = |\mathcal{S}|^{-1} \sum_{i \in \mathcal{S}} \boldsymbol{x}_{i,\tau_i}$

Algorithm 3 $\boldsymbol{p}_{N'} =$ **FEDIHGP** $(\boldsymbol{x}, \boldsymbol{y}^+, N)$: Federated approximation of inverse-Hessian-gradient product

Select $N' \in \{0, \ldots, N-1\}$ and $\mathcal{S}_0 \in \mathcal{S}$ UAR.
for $i \in \mathcal{S}_0$ **in parallel do**
$\quad \boldsymbol{p}_{i,0} = \nabla_y f_i(\boldsymbol{x}, \boldsymbol{y}^+; \xi_{i,0})$
end for
$\boldsymbol{p}_0 = \frac{N}{\ell_{g,1}} |\mathcal{S}_0|^{-1} \sum_{i \in \mathcal{S}_0} \boldsymbol{p}_{i,0}$
if $N' = 0$ **then**
$\quad$ Return $\boldsymbol{p}_{N'}$
end if
Select $\mathcal{S}_1, \ldots, \mathcal{S}_{N'} \in \mathcal{S}$ UAR.
for $n = 1, \ldots, N'$ **do**
$\quad$ **for** $i \in \mathcal{S}_n$ **in parallel do**
$\quad\quad \boldsymbol{p}_{i,n} = \big(\boldsymbol{I} - \frac{1}{\ell_{g,1}} \nabla^2_y g_i(\boldsymbol{x}, \boldsymbol{y}^+; \zeta_{i,n})\big) \boldsymbol{p}_{n-1}$
$\quad$ **end for**
$\quad \boldsymbol{p}_n = |\mathcal{S}_n|^{-1} \sum_{i \in \mathcal{S}_n} \boldsymbol{p}_{i,n}$
end for

the function's input as $(\boldsymbol{x}, \boldsymbol{y}^+)$ (instead of $(\boldsymbol{x}^k, \boldsymbol{y}^{k+1})$) and the output as $\boldsymbol{x}^+$ (instead of $\boldsymbol{x}^{k+1}$). For each client $i \in \mathcal{S}$, FEDOUT uses stochastic approximations of $\bar{\nabla}^{\mathrm{I}} f_i(\boldsymbol{x}, \boldsymbol{y}^+)$ and $\bar{\nabla}^{\mathrm{D}} f_i(\boldsymbol{x}, \boldsymbol{y}^+)$, which we call $\boldsymbol{h}_i^{\mathrm{I}}(\boldsymbol{x}, \boldsymbol{y}^+)$ and $\boldsymbol{h}_i^{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{y}^+)$, respectively. The specific choice of these approximations (see Line 5) is critical and is discussed in detail later in this section. Before that, we explain how each client uses these proxies to form local updates of the outer variable. In each round, starting from a common global model $\boldsymbol{x}_{i,0} = \boldsymbol{x}$, each client i performs τ_i local steps (in parallel):

$$\boldsymbol{x}_{i,\ell+1} = \boldsymbol{x}_{i,\ell} - \alpha_i \boldsymbol{h}_{i,\ell}, \tag{7}$$

and then the server aggregates local models via $\boldsymbol{x}^+ = |\mathcal{S}|^{-1} \sum_{i \in \mathcal{S}} \boldsymbol{x}_{i,\tau_i}$. Here, $\alpha_i \in (0, \alpha]$ is the local stepsize,

$$\begin{aligned} \boldsymbol{h}_{i,\ell} :=& \boldsymbol{h}^{\mathrm{I}}(\boldsymbol{x}, \boldsymbol{y}^+) + \boldsymbol{h}^{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{y}^+) \\ & - \boldsymbol{h}_i^{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{y}^+) + \boldsymbol{h}_i^{\mathrm{D}}(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+), \end{aligned} \tag{8}$$

and, $\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}) := |\mathcal{S}|^{-1} \sum_{i \in \mathcal{S}} \boldsymbol{h}_i(\boldsymbol{x}, \boldsymbol{y}) = |\mathcal{S}|^{-1} \sum_{i \in \mathcal{S}} \boldsymbol{h}_i^{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{y}^+) - \boldsymbol{h}_i^{\mathrm{I}}(\boldsymbol{x}, \boldsymbol{y}^+)$.

The key features of updates (7)–(8) are exploiting past gradients (variance reduction) to account for objective heterogeneity. Indeed, the ideal update in FEDOUT would perform the update $\boldsymbol{x}_{i,\ell+1} = \boldsymbol{x}_{i,\ell} - \alpha_i \big(\boldsymbol{h}^{\mathrm{I}}(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+) + \boldsymbol{h}^{\mathrm{D}}(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+)\big)$ using the global gradient estimates. But this requires each client i to have access to both direct and indirect gradients of all other clients–which it does not, since clients do not communicate between rounds. To overcome this issue, each client i uses global gradient estimates, i.e., $\boldsymbol{h}^{\mathrm{I}}(\boldsymbol{x}, \boldsymbol{y}^+) + \boldsymbol{h}^{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{y}^+)$ from the beginning of each round as a guiding direction in its local update rule. However, since both $\boldsymbol{h}^{\mathrm{D}}$ and $\boldsymbol{h}^{\mathrm{I}}$ are computed at a previous $(\boldsymbol{x}, \boldsymbol{y}^+)$, client i makes a correction by subtracting off the stale direct gradient estimate $\boldsymbol{h}_i^{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{y}^+)$ and adding its own *local* estimate $\boldsymbol{h}_i^{\mathrm{D}}(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+)$. Our local update rule in Step 10 of Algorithm 2 is precisely of this form, i.e. $\boldsymbol{h}_{i,\ell}$ approximates $\boldsymbol{h}^{\mathrm{I}}(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+) + \boldsymbol{h}^{\mathrm{D}}(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+)$ via (8). Note here that the described local correction of FEDOUT only applies to the direct gradient component (the indirect component would require global Hessian information). An alterantive approach leading to LFEDNEST is discussed in Section 2.6.

FEDOUT applied to special nested problems. Algorithm 2 naturally allows the use of other optimizers for min-max & compositional optimization. For example, in the min-max problem (2), the bilevel gradient components are $\nabla^{\mathrm{D}} f_i(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})) = \nabla_{\boldsymbol{x}} f_i(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}))$ and $\nabla^{\mathrm{I}} f_i(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})) = 0$ for all $i \in \mathcal{S}$. Hence, the hypergradient estimate (8) reduces to

$$\boldsymbol{h}_{i,\ell} = \boldsymbol{h}^{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{y}^+) - \boldsymbol{h}_i^{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{y}^+) + \boldsymbol{h}_i^{\mathrm{D}}(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+). \tag{9}$$

More details on these special cases are provided in Appendixes C and E.

Indirect gradient estimation & FEDIHGP. Here, we aim to address one of the key challenges in nested FL: inverse Hessian gradient product. Note from (6b) that the proposed stochastic approximation of the indirect gradient involves in a nonlinear way the *global* Hessian, which is *not* available at the client. To get around this, we use a client sampling strategy and recursive reformulation of (6b) so that $\bar{\nabla}^{\mathrm{I}} f_i(\boldsymbol{x}, \boldsymbol{y})$ can be estimated in an efficient federated manner. In particular, given $N \in \mathbb{N}$, we select $N' \in \{0 \ldots, N-1\}$ and $\mathcal{S}_0, \ldots, \mathcal{S}_{N'} \in \mathcal{S}$ UAR. For all $i \in \mathcal{S}$, we then define

$$\boldsymbol{h}_i^{\mathrm{I}}(\boldsymbol{x}, \boldsymbol{y}) = \nabla^2_{\boldsymbol{xy}} g_i(\boldsymbol{x}, \boldsymbol{y}; \zeta_i) \boldsymbol{p}_{N'}, \tag{10a}$$

where $\boldsymbol{p}_{N'} = |\mathcal{S}_0|^{-1} \widehat{\boldsymbol{H}}_{\boldsymbol{y}} \sum_{i \in \mathcal{S}_0} \nabla_{\boldsymbol{y}} f_i(\boldsymbol{x}, \boldsymbol{y}; \xi_{i,0})$ and $\widehat{\boldsymbol{H}}_{\boldsymbol{y}}$

Algorithm 4 $y^+ = \textsc{FedInn}\,(x, y, \beta)$

$\mathbb{G}_i(\cdot) \leftarrow \nabla_{yy} g_i(x,\cdot)$ (bilevel), $-\nabla_{yy} f_i(x,\cdot)$ (min-max)
$y_{i,0} = y$ and $\beta_i \in (0, \beta]$
for $i \in \mathcal{S}$ **in parallel do**
  $q_i = \mathbb{G}_i(y; \zeta_i)$
end for
$q = |\mathcal{S}|^{-1} \sum_{i\in\mathcal{S}} q_i$
for $i \in \mathcal{S}$ **in parallel do**
  for $\ell = 0, \dots, \tau_i - 1$ **do**
    $q_{i,\ell} = \mathbb{G}_i(y_{i,\ell}; \zeta_{i,\ell}) - \mathbb{G}_i(y; \zeta_{i,\ell}) + q$
    $y_{i,\ell+1} = y_{i,\ell} - \beta_i q_{i,\ell}$
  end for
end for
$y^+ = |\mathcal{S}|^{-1} \sum_{i\in\mathcal{S}} y_{i,\tau_i}$

is the approximate inverse Hessian:

$$\frac{N}{\ell_{g,1}} \prod_{n=1}^{N'} \Big(I - \frac{1}{\ell_{g,1}|\mathcal{S}_n|} \sum_{i=1}^{|\mathcal{S}_n|} \nabla^2_{yy} g_i(x, y; \zeta_{i,n})\Big). \tag{10b}$$

The subroutine FEDIHGP provides a recursive strategy to compute $p_{N'}$ and FEDOUT multiplies $p_{N'}$ with the global Jacobian to drive an indirect gradient estimate. Importantly, *these approximations require only matrix-vector products and vector communications.*

2.5. FEDINN

In FL, each client performs multiple local training steps in isolation on its own data (using for example SGD) before communicating with the server. Due to such local steps, FEDAVG suffers from a *client-drift* effect under objective heterogeneity; that is, the local iterates of each client drift-off towards the minimum of their own local function. In turn, this can lead to convergence to a point different from the global optimum $y^*(x)$ of the inner problem; e.g., see (Mitra et al., 2021). This behavior is particularly undesirable in a nested optimization setting since it directly affects the outer optimization; see, e.g. (Liu et al., 2021, Section 7).

In light of this observation, we build on the recently proposed FEDLIN (Mitra et al., 2021) to solve the inner problem. For each $i \in \mathcal{S}$, let $q_i(x, y)$ denote an unbiased estimate of the gradient $\nabla_{yy} g_i(x, y)$. In each round, starting from a common global model y, each client i performs τ_i local training steps in parallel: $y_{i,\ell+1} = y_{i,\ell} - \beta_i q_{i,\ell}$, where $q_{i,\ell} := q(x, y) - q_i(x, y) + q_i(x_{i,\ell}, y)$, $\beta_i \in (0, \beta]$ is the local inner stepsize, and $q(x, y) := |\mathcal{S}|^{-1} \sum_{i\in\mathcal{S}} q_i(x, y)$. See Algorithm 4.

2.6. Light-FEDNEST for Communication Efficiency

Each FEDNEST epoch k requires $2T + N + 3$ communication rounds as follows: $2T$ rounds for SVRG of FEDINN, N iterations for vector inverse Hessian approximation within FEDIHGP and 3 additional aggregations. Note that, these are vector communications and we fully avoid Hessian communication. In Appendix G, we also propose simplified variants of FEDOUT and FEDIHGP, which are tailored to homogeneous and/or high-dimensional FL settings. These algorithms can then either use *local* Jacobian / inverse Hessian or their approximation, and can use eit SVRG or SGD.

Light-FEDNEST: Specifically, we propose LFEDNEST where each client runs IHGP locally. This reduces the number of rounds to $2T + 1$, saving $N + 2$ rounds (see experiments in Section 4 for performance comparison and Appendix G for further discussion.)

3. Convergence Analysis for FEDNEST

Next, we present convergence results for FEDNEST. All proofs are relegated to Appendixes B–E.

Theorem 3.1. *Suppose Assumptions A and B hold. Further, assume* $\alpha_i^k = \frac{\alpha^k}{\tau_i}, \beta_i^k = \frac{\beta^k}{\tau_i}, \forall i \in \mathcal{S}$, *where*

$$\beta^k = \frac{\bar\beta \alpha^k}{T}, \quad \alpha^k = \min\Big\{\bar\alpha_1, \bar\alpha_2, \bar\alpha_3, \frac{\bar\alpha}{\sqrt{K}}\Big\} \tag{11}$$

for some constants $\bar\alpha_1, \bar\alpha_2, \bar\alpha_3, \bar\alpha, \bar\beta$ *independent of* K. *Then, for any* $T \geq 1$, *the iterates* $\{(x^k, y^k)\}_{k\geq 0}$ *generated by* FEDNEST *satisfy*

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\Big[\big\|\nabla f(x^k)\big\|^2\Big] = \mathcal{O}\Big(\frac{\bar\alpha \max(\sigma_{g,1}^2, \sigma_{g,2}^2, \sigma_f^2)}{\sqrt{K}} + \frac{1}{\min(\bar\alpha_1, \bar\alpha_2, \bar\alpha_3)K} + b^2\Big),$$

where $b := \frac{\ell_{g,1}\ell_{f,1}}{\mu_g}\big(\frac{\kappa_g - 1}{\kappa_g}\big)^N$ *and* N *is the input parameter to* FEDIHGP.

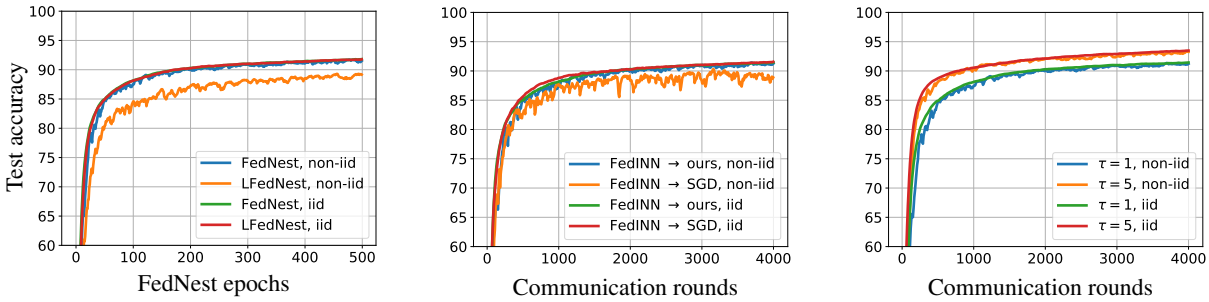
Corollary 3.1 (Bilevel). *Let* $\kappa_g = \ell_{g,1}/\mu_g$. *Under the same conditions as in Theorem 3.1, if we select* $N = \mathcal{O}(\kappa_g \log K)$ *and* $T = \mathcal{O}(\kappa_g^4)$. *Then,*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(x^k)\|^2] = \mathcal{O}\Big(\frac{\kappa_g^3}{K} + \frac{\kappa_g^{2.5}}{\sqrt{K}}\Big). \tag{12}$$

For ϵ*-accurate stationary point, we need* $K = \mathcal{O}(\kappa_g^5 \epsilon^{-2})$.

Above, we choose $N \propto \kappa_g \log K$ to guarantee $b^2 \lesssim 1/\sqrt{K}$. In contrast, we use $T \gtrsim \kappa_g^4$ SVRG epochs. From Section 2.6, this would imply the communication cost is dominated by SVRG epochs N and $\mathcal{O}(\kappa_g^4)$ rounds.

From Corollary 3.1, we remark that FEDNEST matches the guarantees of centralized alternating SGD methods, such as ALSET (Chen et al., 2021a), despite federated setting, i.e. communication challenge, heterogeneity in the client objectives, and device heterogeneity.



(a) Comparison between FEDNEST and LFEDNEST.

(b) SVRG in FEDINN provides better convergence and stability.

(c) Larger τ in FEDOUT provides better performance.

Figure 2: Hyper-representation experiments on a 2-layer MLP and MNIST dataset.

3.1. Reduction to Min-Max FL

We focus on special features of federated min-max problems and customize the general results to yield improved convergence results for this special case. Recall from (2) that $g_i(\boldsymbol{x}, \boldsymbol{y}) = -f_i(\boldsymbol{x}, \boldsymbol{y})$. Then, following Assumption A, $f_i(\boldsymbol{x}, \boldsymbol{y})$ is μ_f strongly concave in $\boldsymbol{y}$ for all $\boldsymbol{x}$.

Corollary 3.2 (Min-Max)**.** *Denote $\kappa_f = \ell_{f,1}/\mu_f$. Assume same conditions as in Theorem 3.1 and $T = \mathcal{O}(\kappa_f)$. Then,*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] = \mathcal{O}\Big(\frac{\kappa_f^2}{K} + \frac{\kappa_f}{\sqrt{K}}\Big). \tag{13}$$

Corrollary 3.2 implies that for the min-max problem, the convergence rate of FEDNEST to the stationary point of f is $\mathcal{O}(1/\sqrt{K})$. Again, we note this matches the convergence rate of non-federated stochastic algorithms (see also Table 1) such as ALSET (Chen et al., 2021a), SGDA (Lin et al., 2020), and SMD (Rafique et al., 2021).

3.2. Reduction to Single-Level FL

Building upon the general results for stochastic nonconvex nested problems, we establish new convergence guarantees for *single-level* stochastic non-convex federated SVRG which is integrated within our FEDOUT.

We make the following assumptions that are counterparts of Assumption A and B.

Assumption C (Lipschitz continuity)**.** *For all $i \in [m]$, $\nabla f_i(\boldsymbol{x})$ is $\ell_{f,1}$-Lipschitz continuous.*

Assumption D (Stochastic samples)**.** *For all $i \in [m]$, $\nabla f_i(\boldsymbol{x};\xi)$ is an unbiased estimator of $\nabla f_i(\boldsymbol{x})$ and its variance is bounded, i.e., $\mathbb{E}_\xi[\|\nabla f_i(\boldsymbol{x};\xi) - \nabla f_i(\boldsymbol{x})\|^2] \le \sigma_f^2$.*

Theorem 3.2 (Single-Level)**.** *Suppose Assumptions C and D hold. Further, assume $\alpha_i^k = \frac{\alpha^k}{\tau_i}$, $\forall i \in \mathcal{S}$, where*

$$\alpha^k = \bar{\alpha}\min\Big\{\frac{1}{L_f}, \frac{1}{\sqrt{K}\sigma_f}\Big\} \tag{14}$$

for some constants $\bar{\alpha}$. Then,

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] = \mathcal{O}\Big(\frac{1}{K} + \frac{\sigma_f}{\sqrt{K}}\Big). \tag{15}$$

Theorem 3.2 extends very recent results by (Mitra et al., 2021) from the stochastic strongly convex to the stochastic nonconvex setting.

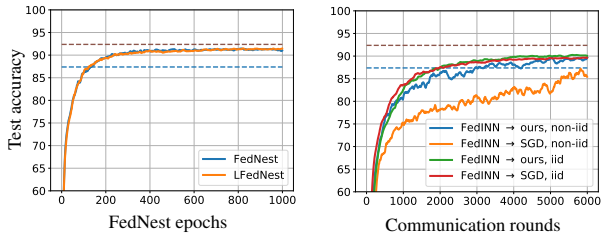
4. Numerical Result

In this section, we first numerically investigate the impact of several attributes of our algorithms on a simple hyper-representation problem similar to (Franceschi et al., 2018). Then, we test our algorithm on a hyper-parameter optimization problem for loss function tuning in imbalanced training following (Li et al., 2021).

4.1. Hyper-representation learning

Modern approaches in meta learning (ML) such as MAML (Finn et al., 2017) and reptile (Nichol & Schulman, 2018) learn representations (that are shared across all tasks) in a bilevel manner. Similarly, the hyper-representation problem optimizes a classification model in a two-phased process. The outer objective optimizes the model backbone to obtain better feature representation on validation data. The inner problem optimizes a header for downstream classification tasks on training data. In this experiment, we use a 2-layer multilayer perceptron (MLP) with 200 hidden units. The outer problem optimizes the hidden layer with 157,000 parameters, and the inner problem optimizes the output layer with 2,010 parameters. We study both i.i.d and non-i.i.d. ways of partitioning the MNIST data exactly following FEDAVG (McMahan et al., 2017), and split each client's data evenly to train and validation datasets. Thus, each client has 300 train and 300 validation samples.

Fig 2 demonstrates the impact on test accuracy of several im-



(a) FEDNEST achieves similar performance as centralized bilevel loss tuning.

(b) SVRG in FEDINN provides better convergence and stability especially in non-iid setup.

Figure 3: Imbalanced loss tuning on a 3-layer MLP and imbalanced MNIST dataset. The brown dashed line is the test accuracy on non-federated bilevel optimization, and the blue dashed line is the test accuracy without loss tuning.

portant components of FEDNEST. Fig 2a compares between FEDNEST and LFEDNEST. Both algorithms perform well on the i.i.d. setup, while on the non-i.i.d. setup, FEDNEST achieves i.i.d. performance, significantly outperforming LFEDNEST. These findings are in line with our discussions in Sec. 2.6 and Fig. 2a. LFEDNEST saves on communication rounds compared to FEDNEST and performs well on homogeneous clients. But, for heterogeneous clients, the isolation of local Hessian in LFEDNEST (see Algo. 5) degrades the test performance. Next, Fig. 2b demonstrates the importance of SVRG in FEDINN algorithm for heterogeneous data (as predicted by our theoretical considerations in Sec. 2.5). Specifically, we compare FEDINN with SVRG (Algo 4) with an alternative SGD-based implementation. Finally, Fig. 2c elucidates the role of local epoch τ in FEDOUT: larger τ saves on communication and improves test performance by providing faster convergence speed.

4.2. Loss function tuning on imbalanced dataset

In this section, we use bilevel optimization to tune a loss function for learning an imbalanced MNIST dataset, following the optimization formulation in (Li et al., 2021) (specifically, we tune the so-called VS-loss(Kini et al., 2021)). Unlike (Li et al., 2021), we experiment on a federated setting. Specifically, we first create a long-tail imbalanced MNIST dataset by exponentially decaying the examples for each class (e.g. class 0 has 6,000 samples, class 1 has 3,597 samples and finally, class 9 has only 60 samples). We assign data to 100 clients following again FEDAVG(McMahan et al., 2017) on both i.i.d. and non-i.i.d. setups. Different from the hyper-representation experiment, we employ 80%-20% train-validation on each client and use a 3-layer MLP model with 200, 100 hidden units, respectively. It is worth noting that, in this problem, the outer objective f (aka validation cost) only depends on the hyperparameter $\boldsymbol{x}$ through the optimal model parameters $\boldsymbol{y}^*(\boldsymbol{x})$; thus, the direct gradient $\nabla^{\mathrm{D}} f_i(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}))$ is zero.

Fig. 3 displays test accuracy vs epochs/rounds for our FL bilevel algorithms. The horizontal dashed lines serve as centralized baselines: brown depicts accuracy reached by non-federated bilevel optimization, and, blue depicts accuracy without any loss tuning. Compared to these, Fig. 3a shows that FEDNEST achieves near centralized performance. In Fig. 3b, we investigate the key role of SVRG in FEDINN by comparing it with an possible alternative implementation that uses SGD-type updates. The figure confirms our discussion in Sec. 2.5: SVRG offers significant performance gains that are pronounced by client heterogeneity (e.g. compare blue to orange curve).

5. Related Work

To the best of our knowledge, FEDNEST is the first approach that can optimize a general bilevel optimization problem in a FL setting. Related work falls into two categories: federated learning and bilevel optimization.

Federated learning. FEDAVG was first introduced by McMahan et al. (2017), who showed it can dramatically reduce communication costs. For identical clients, FEDAVG coincides with local SGD (Zinkevich et al., 2010) which has been analyzed by many works (Stich, 2019; Yu et al., 2019; Wang & Joshi, 2018; Stich & Karimireddy, 2019; Basu et al., 2019). Recently, many variants of FEDAVG have been proposed to tackle issues such as convergence and client drift. Examples include FEDPROX (Li et al., 2020b), SCAFFOLD (Karimireddy et al., 2019), FEDSPLIT (Pathak & Wainwright, 2020), FEDNOVA (Wang et al., 2020), and, the most closely relevant to us FEDLIN (Mitra et al., 2021).

Bilevel optimization. This class of problems was first introduced by (Bracken & McGill, 1973), and since then, different types of approaches have been proposed. See (Sinha et al., 2017; Liu et al., 2021) for surveys. Earlier works in (Aiyoshi & Shimizu, 1984; Edmunds & Bard, 1991; Lv et al., 2007) reduced the bilevel problem to a single-level optimization problem. However, the reduced problem is still difficult to solve due to for example a large number of constraints. Recently, more efficient gradient-based algorithms have been proposed by estimating the hypergradient of $\nabla f(\boldsymbol{x})$ through iterative updates (Maclaurin et al., 2015; Franceschi et al., 2017; Finn et al., 2017; Grazzi et al., 2020; Domke, 2012; Pedregosa, 2016; Grazzi et al., 2020). Theoretically, bilevel optimization has been studied via both asymptotic (Franceschi et al., 2018; Shaban et al., 2019; Liu et al., 2020; Li et al., 2020a) and finite-time (non-asymptotic) analysis (Ghadimi & Wang, 2018; Hong et al., 2020; Ji et al., 2021; Chen et al., 2021a).

A more in-depth discussion of related work is given in Appendix A. We summarize the complexities of different methods for FL/non-FL bilevel optimization in Tables 1 and 2.

References

Aiyoshi, E. and Shimizu, K. A solution method for the static constrained stackelberg problem via penalty method. *IEEE Transactions on Automatic Control*, 29(12):1111–1114, 1984.

Al-Khayyal, F. A., Horst, R., and Pardalos, P. M. Global optimization of concave functions subject to quadratic constraints: an application in nonlinear bilevel programming. *Annals of Operations Research*, 34(1):125–147, 1992.

Arora, S., Du, S., Kakade, S., Luo, Y., and Saunshi, N. Provable representation learning for imitation learning via bi-level optimization. In *International Conference on Machine Learning*, pp. 367–376. PMLR, 2020.

Basu, D., Data, D., Karakus, C., and Diggavi, S. Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations. In *Advances in Neural Information Processing Systems*, pp. 14668–14679, 2019.

Bertinetto, L., Henriques, J. F., Torr, P. H., and Vedaldi, A. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018.

Bracken, J. and McGill, J. T. Mathematical programs with optimization problems in the constraints. *Operations Research*, 21(1):37–44, 1973.

Chen, T., Sun, Y., and Yin, W. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Advances in Neural Information Processing Systems*, 34, 2021a.

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

Dai, B., He, N., Pan, Y., Boots, B., and Song, L. Learning from conditional distributions via dual embeddings. In *Artificial Intelligence and Statistics*, pp. 1458–1467. PMLR, 2017.

Daskalakis, C. and Panageas, I. The limit points of (optimistic) gradient descent in min-max optimization. *arXiv preprint arXiv:1807.03907*, 2018.

Domke, J. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, pp. 318–326. PMLR, 2012.

Edmunds, T. A. and Bard, J. F. Algorithms for nonlinear bilevel mathematical programs. *IEEE transactions on Systems, Man, and Cybernetics*, 21(1):83–89, 1991.

Feurer, M. and Hutter, F. Hyperparameter optimization. In *Automated machine learning*, pp. 3–33. Springer, Cham, 2019.

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

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

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

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

Ghadimi, S., Ruszczynski, A., and Wang, M. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM Journal on Optimization*, 30(1):960–979, 2020.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

Grazzi, R., Franceschi, L., Pontil, M., and Salzo, S. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pp. 3748–3758. PMLR, 2020.

Guo, Z., Xu, Y., Yin, W., Jin, R., and Yang, T. On stochastic moving-average estimators for non-convex optimization. *arXiv preprint arXiv:2104.14840*, 2021.

Hansen, P., Jaumard, B., and Savard, G. New branch-and-bound rules for linear bilevel programming. *SIAM Journal on scientific and Statistical Computing*, 13(5):1194–1217, 1992.

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

Hsu, T.-M. H., Qi, H., and Brown, M. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

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

Ji, K. and Liang, Y. Lower bounds and accelerated algorithms for bilevel optimization. *ArXiv*, abs/2102.03926, 2021.

Ji, K., Yang, J., and Liang, Y. Provably faster algorithms for bilevel optimization and applications to meta-learning. *ArXiv*, abs/2010.07962, 2020a.

Ji, K., Yang, J., and Liang, Y. Theoretical convergence of multi-step model-agnostic meta-learning. *arXiv preprint arXiv:2002.07836*, 2020b.

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

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. SCAFFOLD: Stochastic controlled averaging for on-device federated learning. *arXiv preprint arXiv:1910.06378*, 2019.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Khaled, A., Mishchenko, K., and Richtárik, P. First analysis of local GD on heterogeneous data. *arXiv preprint arXiv:1909.04715*, 2019.

Khanduri, P., Zeng, S., Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *arXiv preprint arXiv:2102.07367*, 2021.

Kini, G. R., Paraskevas, O., Oymak, S., and Thrampoulidis, C. Label-imbalanced and group-sensitive classification under overparameterization. *arXiv preprint arXiv:2103.01550*, 2021.

Konečnỳ, J., McMahan, H. B., Ramage, D., and Richtárik, P. Federated optimization: Distributed machine learning for on-device intelligence. *International Conference on Learning Representations*, 2018.

Li, J., Gu, B., and Huang, H. Improved bilevel model: Fast and optimal algorithm with theoretical guarantee. *arXiv preprint arXiv:2009.00690*, 2020a.

Li, M., Zhang, X., Thrampoulidis, C., Chen, J., and Oymak, S. Autobalance: Optimized loss functions for imbalanced data. *Advances in Neural Information Processing Systems*, 34, 2021.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020b.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of FedAvg on non-IID data. *arXiv preprint arXiv:1907.02189*, 2019.

Lin, T., Jin, C., and Jordan, M. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pp. 6083–6093. PMLR, 2020.

Liu, H., Simonyan, K., and Yang, Y. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

Liu, R., Mu, P., Yuan, X., Zeng, S., and Zhang, J. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *International Conference on Machine Learning*, pp. 6305–6315. PMLR, 2020.

Liu, R., Gao, J., Zhang, J., Meng, D., and Lin, Z. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *arXiv preprint arXiv:2101.11517*, 2021.

Lorraine, J., Vicol, P., and Duvenaud, D. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics*, pp. 1540–1552. PMLR, 2020.

Lv, Y., Hu, T., Wang, G., and Wan, Z. A penalty function method based on kuhn–tucker condition for solving linear bilevel programming. *Applied Mathematics and Computation*, 188(1):808–813, 2007.

Maclaurin, D., Duvenaud, D., and Adams, R. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pp. 2113–2122. PMLR, 2015.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, pp. 1273–1282, 2017. URL http://proceedings.mlr.press/v54/mcmahan17a.html.

Mitra, A., Jaafar, R., Pappas, G., and Hassani, H. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. *Advances in Neural Information Processing Systems*, 34, 2021.

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

Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pp. 1497–1507. PMLR, 2020.

Moore, G. M. *Bilevel programming algorithms for machine learning model selection*. Rensselaer Polytechnic Institute, 2010.

Nichol, A. and Schulman, J. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2(3):4, 2018.

Pathak, R. and Wainwright, M. J. Fedsplit: an algorithmic framework for fast federated optimization. *Advances in Neural Information Processing Systems*, 33:7057–7066, 2020.

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

Rafique, H., Liu, M., Lin, Q., and Yang, T. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, pp. 1–35, 2021.

Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečnỳ, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. In *International Conference on Learning Representations*, 2020.

Shaban, A., Cheng, C.-A., Hatch, N., and Boots, B. Truncated back-propagation for bilevel optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1723–1732. PMLR, 2019.

Shi, C., Lu, J., and Zhang, G. An extended kuhn–tucker approach for linear bilevel programming. *Applied Mathematics and Computation*, 162(1):51–63, 2005.

Sinha, A., Malo, P., and Deb, K. A review on bilevel optimization: from classical to evolutionary approaches and applications. *IEEE Transactions on Evolutionary Computation*, 22(2):276–295, 2017.

Stich, S. U. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1g2JnRcFX.

Stich, S. U. and Karimireddy, S. P. The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019.

Thekumparampil, K. K., Jain, P., Netrapalli, P., and Oh, S. Efficient algorithms for smooth minimax optimization. *arXiv preprint arXiv:1907.01543*, 2019.

Wang, J. and Joshi, G. Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 2020.

Wang, M., Fang, E. X., and Liu, H. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2):419–449, 2017.

Wang, S., Tuor, T., Salonidis, T., Leung, K. K., Makaya, C., He, T., and Chan, K. Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*, 37(6): 1205–1221, 2019.

Yu, H., Yang, S., and Zhu, S. Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5693–5700, 2019.

Zinkevich, M., Weimer, M., Li, L., and Smola, A. J. Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, pp. 2595–2603, 2010.

APPENDIX
FEDNEST: Federated Bilevel Optimization

The appendix is organized as follows: Section A provides related work. We give all details for the proof of the main theorems in Sections B, C, D, and E, for federated bilevel, min-max, single-level, and compositional optimziation, respectively. In Section F, we state a few auxiliary technical lemmas. Finally, Section G discusses LFEDNEST algorithm.

A. Related Work

Bilevel Optimization. A broad collection of algorithms have been proposed to solve bilevel nonlinear programming problems (Sinha et al., 2017; Liu et al., 2021). Earlier works in (Aiyoshi & Shimizu, 1984; Edmunds & Bard, 1991; Al-Khayyal et al., 1992; Hansen et al., 1992; Shi et al., 2005; Lv et al., 2007; Moore, 2010) reduced the bilevel problem to a single-level optimization problem using the Karush-Kuhn-Tucker (KKT) conditions or penalty function methods. However, the reduced problem is still difficult to solve (Sinha et al., 2017). In comparison, gradient-based approaches are more attractive due to their simplicity and effectiveness. This type of approach estimates the hypergradient $\nabla\varphi(\boldsymbol{x})$ for iterative updates, and are generally divided into approximate implicit differentiation (AID)- and iterative differentiation (ITD)-based categories. ITD-based approaches (Maclaurin et al., 2015; Franceschi et al., 2017; Finn et al., 2017; Grazzi et al., 2020) estimate the hypergradient $\nabla\varphi(\boldsymbol{x})$ in either a reverse (automatic differentiation) or forward manner. AID-based approaches (Domke, 2012; Pedregosa, 2016; Grazzi et al., 2020; Ji et al., 2021) estimate the hypergradient via implicit differentiation which involves solving a linear system.

Theoretically, bilevel optimization has been studied via both asymptotic and finite-time (non-asymptotic) analysis (Franceschi et al., 2018; Liu et al., 2020; Li et al., 2020a; Shaban et al., 2019; Ghadimi & Wang, 2018; Ji et al., 2021; Hong et al., 2020). In particular, (Franceschi et al., 2018) provided the asymptotic convergence of a backpropagation-based approach as one of ITD-based algorithms by assuming the inner-level problem is strongly convex. (Shaban et al., 2019) gave a similar analysis for a *truncated* backpropagation approach. Finite-time complexity analysis for bilevel optimization has also been explored. In particular, (Ghadimi & Wang, 2018) provided a finite-time convergence analysis for an AID-based algorithm under two different loss geometries, where $\varphi(\cdot)$ is strongly convex, convex or nonconvex, and $g(\boldsymbol{x},\cdot)$ is strongly convex. (Ji et al., 2021) provided an improved finite-time analysis for AID- and ITD-based algorithms under the nonconvex-strongly-convex geometry. (Ji & Liang, 2021) provided the first-known lower bounds on complexity as well as tighter upper bounds. When the objective functions can be expressed in an expected or finite-time form, (Ghadimi & Wang, 2018; Ji et al., 2021; Hong et al., 2020) developed stochastic bilevel algorithms and provided the finite-time analysis. (Chen et al., 2021a) provided a tighter analysis of SGD for stochastic bilevel problems. (Chen et al., 2021b; Guo et al., 2021; Khanduri et al., 2021; Ji et al., 2020a; Huang & Huang, 2021) studied accelerated SGD, momentum, and adaptive-type bilevel optimization methods. More results can be found in the recent review paper (Liu et al., 2021) and references therein.

Federated Learning involves learning a centralized model from distributed client data. Although this centralized model benefits from all client data, it raises several types of issues such as generalization, fairness, communication efficiency, and privacy (Mohri et al., 2019; Stich, 2019; Yu et al., 2019; Wang & Joshi, 2018; Stich & Karimireddy, 2019; Basu et al., 2019). FEDAVG McMahan et al. (2017) can tackle some of these issues such as high communication costs. Many variants of FEDAVG have been proposed to tackle issues such as convergence and *client drift*. Examples include adding a regularization term in the client objectives towards the broadcast model (Li et al., 2018), proximal splitting (Pathak & Wainwright, 2020; Mitra et al., 2021), variance reduction (Karimireddy et al., 2019; Mitra et al., 2021) and adaptive updates (Reddi et al., 2020). When clients are homogeneous, FEDAVG is closely related to local SGD (Zinkevich et al., 2010), which has been analyzed by many works (Stich, 2019; Yu et al., 2019; Wang & Joshi, 2018; Stich & Karimireddy, 2019; Basu et al., 2019). In order to analyze FEDAVG in heterogeneous settings, (Li et al., 2018; Wang et al., 2019; Khaled et al., 2019; Li et al., 2019) derive convergence rates depending on the amount of heterogeneity. They showed that the convergence rate of FEDAVG gets worse with client heterogeneity. By using control variates to reduce client drift, the SCAFFOLD method (Karimireddy et al., 2019) achieves convergence rates that are independent of the amount of heterogeneity. FEDNOVA (Wang et al., 2020) and FEDLIN (Mitra et al., 2021) provided the convegence of their methods despite arbitrary local objective and systems heterogeneity. In particular, (Mitra et al., 2021) showed that FEDLIN guarantees linear convergence to the global minimum of deterministic objective, despite arbitrary objective and systems heterogeneity.

B. Proof for Federated Bilevel Optimization

Throughout the proof, we will use $\mathcal{F}_{i,\ell}^{k,t}$ to denote the filtration that captures all the randomness up to the ℓ-th local step of client i in inner round t and outer round k. With a slight abuse of notation, $\mathcal{F}_{i,-1}^{k,t}$ is to be interpreted as $\mathcal{F}^{k,t}, \forall i \in \mathcal{S}$. For simplicity, we remove subscripts k and t from the definition of stepsize and model parameters. For example, $\boldsymbol{x}$ and $\boldsymbol{x}^+$ denote $\boldsymbol{x}^k$ and $\boldsymbol{x}^{k+1}$, respectively. We further set

$$\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+) := \mathbb{E}\left[\boldsymbol{h}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+)|\mathcal{F}_{i,\ell-1}\right]. \tag{16}$$

The following lemma extends (Ghadimi & Wang, 2018, Lemma 2.2) and (Chen et al., 2021a, Lemma 2) to the finite-sum problem (1). Proofs follow similarly by applying their analysis to the inner & outer functions $(f_i, g_i), \forall i \in \mathcal{S}$.

Lemma B.1. *Under Assumptions A and B, the following holds for all $\boldsymbol{x}_1$, $\boldsymbol{x}_2$:*

$$\|\nabla f(\boldsymbol{x}_1) - \nabla f(\boldsymbol{x}_2)\| \le L_f \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|, \tag{17a}$$
$$\|\boldsymbol{y}^*(\boldsymbol{x}_1) - \boldsymbol{y}^*(\boldsymbol{x}_2)\| \le L_y \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|, \tag{17b}$$
$$\|\nabla \boldsymbol{y}^*(\boldsymbol{x}_1) - \nabla \boldsymbol{y}^*(\boldsymbol{x}_2)\| \le L_{yx} \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|, \tag{17c}$$

Also, for all $i \in \mathcal{S}$ and all $\boldsymbol{x}$, $\boldsymbol{y}$, we have

$$\|\bar{\nabla} f_i(\boldsymbol{x}, \boldsymbol{y}) - \bar{\nabla} f_i(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}))\| \le M_f \|\boldsymbol{y}^*(\boldsymbol{x}) - \boldsymbol{y}\|, \tag{17d}$$
$$\mathbb{E}\left[\|\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}) - \boldsymbol{h}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y})\|^2\right] \le \tilde{\sigma}_f^2, \tag{17e}$$
$$\mathbb{E}\left[\|\boldsymbol{h}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+)\|^2 |\mathcal{F}_{i,\ell-1}\right] \le \tilde{D}_f^2. \tag{17f}$$

Here,

$$L_{yx} := \frac{\ell_{g,2} + \ell_{g,2} L_y}{\mu_g} + \frac{\ell_{g,1}(\ell_{g,2} + \ell_{g,2} L_y)}{\mu_g^2}, \quad L_y := \frac{\ell_{g,1}}{\mu_g},$$
$$M_f := \ell_{f,1} + \frac{\ell_{g,1}\ell_{f,1}}{\mu_g} + \frac{\ell_{f,0}}{\mu_g}\left(\ell_{g,2} + \frac{\ell_{g,1}\ell_{g,2}}{\mu_g}\right),$$
$$L_f := \ell_{f,1} + \frac{\ell_{g,1}(\ell_{f,1} + L_f)}{\mu_g} + \frac{\ell_{f,0}}{\mu_g}\left(\ell_{g,2} + \frac{\ell_{g,1}\ell_{g,2}}{\mu_g}\right),$$
$$\tilde{\sigma}_f^2 := \sigma_f^2 + \frac{3}{\mu_g^2}\left[(\sigma_f^2 + \ell_{f,0}^2)(\sigma_{g,2}^2 + 2\ell_{g,1}^2) + \sigma_f^2 \ell_{g,1}^2\right],$$
$$\tilde{D}_f^2 := \left(\ell_{f,0} + \frac{\ell_{g,1}}{\mu_g}\ell_{f,1} + \ell_{g,1}\ell_{f,1}\frac{1}{\mu_g}\right)^2 + \tilde{\sigma}_f^2,$$

where the other constants are provided in Assumptions A and B.

B.1. Descent of Outer Objective

The following lemma characterizes the descent of the outer objective.

Lemma B.2 (Descent Lemma)**.** *Suppose Assumptions A-B hold. Further, assume $\tau_i \ge 1$ and $\alpha_i = \alpha/\tau_i, \forall i \in \mathcal{S}$ for some positive constant α. Then,* FEDOUT *guarantees:*

$$\begin{aligned}\mathbb{E}\left[f(\boldsymbol{x}^+)\right] - \mathbb{E}\left[f(\boldsymbol{x})\right] \le -\frac{\alpha}{2}(1 - \alpha L_f)\mathbb{E}\left[\|\frac{1}{m}\sum_{i=1}^m \frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1} \bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+)\|^2\right] - \frac{\alpha}{2}\mathbb{E}\left[\|\nabla f(\boldsymbol{x})\|^2\right] \\ + \frac{3\alpha}{2}\left(b^2 + M_f^2 \mathbb{E}\left[\|\boldsymbol{y}^+ - \boldsymbol{y}^*(\boldsymbol{x})\|^2\right] + \frac{M_f^2}{m}\sum_{i=1}^m \frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\mathbb{E}\left[\|\boldsymbol{x}_{i,\ell} - \boldsymbol{x}\|^2\right]\right) + \frac{\alpha^2 L_f}{2}\tilde{\sigma}_f^2.\end{aligned} \tag{18}$$

Proof. It follows from Algorithm 2 that $\boldsymbol{x}_{i,0} = \boldsymbol{x}, \forall i \in \mathcal{S}$, and

$$\boldsymbol{x}^+ = \boldsymbol{x} - \frac{1}{m}\sum_{i=1}^m \alpha_i \sum_{\ell=0}^{\tau_i-1} \boldsymbol{h}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+). \tag{19}$$

Now, using the Lipschitz property of ∇f in Lemma B.1, we have

$$\begin{aligned}\mathbb{E}\left[f(\boldsymbol{x}^+)\right]-\mathbb{E}\left[f(\boldsymbol{x})\right]&\leq\mathbb{E}\left[\langle\boldsymbol{x}^+-\boldsymbol{x},\nabla f(\boldsymbol{x})\rangle\right]+\frac{L_f}{2}\mathbb{E}\left[\|\boldsymbol{x}^+-\boldsymbol{x}\|^2\right]\\&=-\mathbb{E}\Big[\big\langle\frac{1}{m}\sum_{i=1}^m\alpha_i\sum_{\ell=0}^{\tau_i-1}\boldsymbol{h}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+),\nabla f(\boldsymbol{x})\big\rangle\Big]\\&+\frac{L_f}{2}\mathbb{E}\Big[\big\|\frac{1}{m}\sum_{i=1}^m\alpha_i\sum_{\ell=0}^{\tau_i-1}\boldsymbol{h}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+)\big\|^2\Big].\end{aligned}\tag{20}$$

In the following, we bound each term in (20).

For the first term on the right hand side (RHS) in (20), we have

$$\begin{aligned}-\mathbb{E}\Big[\big\langle\frac{1}{m}\sum_{i=1}^m\alpha_i\sum_{\ell=0}^{\tau_i-1}\boldsymbol{h}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+),\nabla f(\boldsymbol{x})\big\rangle\Big]&=-\mathbb{E}\Big[\big\langle\frac{1}{m}\sum_{i=1}^m\alpha_i\sum_{\ell=0}^{\tau_i-1}\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+),\nabla f(\boldsymbol{x})\big\rangle\Big]\\&=-\frac{\alpha}{2}\mathbb{E}\Big[\big\|\frac{1}{m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+)\|^2\Big]-\frac{\alpha}{2}\mathbb{E}\left[\|\nabla f(\boldsymbol{x})\|^2\right]\\&+\frac{\alpha}{2}\mathbb{E}\left[\|\frac{1}{m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+)-\nabla f(\boldsymbol{x})\|^2\right],\end{aligned}\tag{21}$$

where the first equality uses the fact that $\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+)=\mathbb{E}\left[\boldsymbol{h}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+)|\mathcal{F}_{i,\ell-1}\right]$; and the second equality follows from our assumption $\alpha_i=\alpha/\tau_i,\forall i\in\mathcal{S}$.

Next, we bound the last term in (21). Note that

$$\begin{aligned}\left\|\frac{1}{m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+)-\nabla f(\boldsymbol{x})\right\|^2&=\Big\|\frac{1}{m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}(\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+)-\bar{\nabla}f_i(\boldsymbol{x},\boldsymbol{y}^+))\\&+\frac{1}{m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\bar{\nabla}f_i(\boldsymbol{x},\boldsymbol{y}^+)-\nabla f(\boldsymbol{x})\Big\|^2\\&\leq3\left\|\frac{1}{m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}(\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+)-\bar{\nabla}f_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+))\right\|^2\\&+3\left\|\frac{1}{m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}(\bar{\nabla}f_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+)-\bar{\nabla}f_i(\boldsymbol{x},\boldsymbol{y}^+))\right\|^2\\&+3\left\|\bar{\nabla}f(\boldsymbol{x},\boldsymbol{y}^+)-\nabla f(\boldsymbol{x})\right\|^2,\end{aligned}\tag{22}$$

where the inequality uses Lemma F.1.

Hence,

$$\begin{aligned}&\mathbb{E}\left[\|\frac{1}{m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+)-\nabla f(\boldsymbol{x})\|^2\right]\\&\leq3b^2+\frac{3M_f^2}{m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\mathbb{E}\left[\|\boldsymbol{x}_{i,\ell}-\boldsymbol{x}\|^2\right]+3M_f^2\mathbb{E}\left[\|\boldsymbol{y}^+-\boldsymbol{y}^*(\boldsymbol{x})\|^2\right],\end{aligned}\tag{23}$$

where the inequality uses Lemmas B.1 and F.4.

Substituting (23) in (21) yields

$$
\begin{aligned}
&-\mathbb{E}\Big[\big\langle\frac{1}{m}\sum_{i=1}^{m}\alpha_i\sum_{\ell=0}^{\tau_i-1}\boldsymbol{h}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+),\nabla f(\boldsymbol{x})\big\rangle\Big]\\
&\le-\frac{\alpha}{2}\mathbb{E}\left[\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+)\|^2\right]-\frac{\alpha}{2}\mathbb{E}\left[\|\nabla f(\boldsymbol{x})\|^2\right]\\
&+\frac{3\alpha}{2}\Big(b^2+M_f^2\mathbb{E}\left[\|\boldsymbol{y}^+-\boldsymbol{y}^*(\boldsymbol{x})\|^2\right]+\frac{M_f^2}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\mathbb{E}\left[\|\boldsymbol{x}_{i,\ell}-\boldsymbol{x}\|^2\right]\Big).
\end{aligned}
\tag{24}
$$

Next, we bound the second term in (20). Observe that

$$
\begin{aligned}
&\mathbb{E}\left[\Big\|\frac{1}{m}\sum_{i=1}^{m}\alpha_i\sum_{\ell=0}^{\tau_i-1}\boldsymbol{h}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+)\Big\|^2\right]\\
&=\alpha^2\mathbb{E}\left[\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\big(\boldsymbol{h}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+)-\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+)+\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+)\big)\Big\|^2\right]\\
&\le\alpha^2\mathbb{E}\left[\Big\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+)\Big\|^2\right]+\alpha^2\tilde{\sigma}_f^2,
\end{aligned}
\tag{25}
$$

where the inequality follows from Lemmas F.3 and B.1.

Plugging (25) and (24) into (20) completes the proof. □

B.2. Error of FEDINN

The following lemma establishes the progress of FEDINN. It should be mentioned that the assumptions on $\beta_i, \forall i \in \mathcal{S}$ are identical to the ones listed in (Mitra et al., 2021, Theorem 4).

Lemma B.3 (Error of FEDINN)**.** *Suppose Assumptions A-B hold. Further, assume $\tau_i \ge 1$ and $\alpha_i = \alpha/\tau_i, \beta_i = \beta/\tau_i, \forall i \in \mathcal{S}$, where $\beta < \min\big(1/(6\ell_{g,1}), 1\big)$ and α is some positive constant. Then,* FEDINN *guarantees:*

$$
\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{y}^+-\boldsymbol{y}^\star(\boldsymbol{x})\right\|^2\right]&\le\left(1-\frac{\beta\mu_g}{2}\right)^T\mathbb{E}\left[\left\|\boldsymbol{y}-\boldsymbol{y}^\star(\boldsymbol{x})\right\|^2\right]+25T\beta^2\sigma_{g,1}^2,\quad\text{and}\\
\mathbb{E}\left[\left\|\boldsymbol{y}^+-\boldsymbol{y}^\star(\boldsymbol{x}^+)\right\|^2\right]&\le a_1(\alpha)\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+)\right\|^2\right]\\
&\quad+a_2(\alpha)\mathbb{E}\left[\left\|\boldsymbol{y}^+-\boldsymbol{y}^\star(\boldsymbol{x})\right\|^2\right]+a_3(\alpha)\tilde{\sigma}_f^2.
\end{aligned}
$$

Here,

$$
\begin{aligned}
a_1(\alpha)&:=L_y\alpha^2+\frac{L_y\alpha}{4M_f}+\frac{L_{yx}\alpha^2}{2\eta},\\
a_2(\alpha)&:=1+4M_fL_y\alpha+\frac{\eta L_{yx}\tilde{D}_f^2\alpha^2}{2},\\
a_3(\alpha)&:=\alpha^2L_y^2+\frac{L_{yx}\alpha^2}{2\eta}.
\end{aligned}
\tag{26}
$$

for any $\eta > 0$.

Proof. Note that

$$
\begin{aligned}
\mathbb{E}[\|\boldsymbol{y}^+-\boldsymbol{y}^*(\boldsymbol{x}^+)\|^2]=&\mathbb{E}[\|\boldsymbol{y}^+-\boldsymbol{y}^*(\boldsymbol{x})\|^2]+\mathbb{E}[\|\boldsymbol{y}^*(\boldsymbol{x}^+)-\boldsymbol{y}^*(\boldsymbol{x})\|^2]\\
&+2\mathbb{E}[\langle\boldsymbol{y}^+-\boldsymbol{y}^*(\boldsymbol{x}),\boldsymbol{y}^*(\boldsymbol{x})-\boldsymbol{y}^*(\boldsymbol{x}^+)\rangle].
\end{aligned}
\tag{27}
$$

Next, we upper bound each term in (27).

Bounding first term in (27):

From (Mitra et al., 2021, Theorem 4), for all $t \in \{0, \ldots, T-1\}$, we obtain

$$\mathbb{E}[\|\boldsymbol{y}^{t+1} - \boldsymbol{y}^\star(\boldsymbol{x})\|^2] \leq \left(1 - \frac{\beta\mu_g}{2}\right) \mathbb{E}[\|\boldsymbol{y}^t - \boldsymbol{y}^\star(\boldsymbol{x})\|^2] + 25\beta^2\sigma_{g,1}^2,$$

which together with our setting $\boldsymbol{y}^+ = \boldsymbol{y}^T$ gives

$$\mathbb{E}[\|\boldsymbol{y}^+ - \boldsymbol{y}^*(\boldsymbol{x})\|^2] \leq \left(1 - \frac{\beta\mu_g}{2}\right)^T \mathbb{E}[\|\boldsymbol{y} - \boldsymbol{y}(\boldsymbol{x}^*)\|^2] + 25T\beta^2\sigma_{g,1}^2. \tag{28}$$

Bounding second term in (27):

By similar steps as in (25), we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{y}^*(\boldsymbol{x}^+) - \boldsymbol{y}^*(\boldsymbol{x})\|^2] &\leq L_y^2\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m \alpha_i \sum_{\ell=0}^{\tau_i-1} \boldsymbol{h}_i(\boldsymbol{x}, \boldsymbol{y}^+)\right\|^2\right] \\ &\leq L_y^2\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m \alpha_i \sum_{\ell=0}^{\tau_i-1} \bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+)\right\|^2\right] + \alpha^2 L_y^2\tilde{\sigma}_f,\end{aligned} \tag{29}$$

where the inequalities are obtained from Lemmas B.1 and F.3.

Bounding third term in (27):

Observe that

$$\begin{aligned}\mathbb{E}[\langle \boldsymbol{y}^+ - \boldsymbol{y}^*(\boldsymbol{x}), \boldsymbol{y}^*(\boldsymbol{x}) - \boldsymbol{y}^*(\boldsymbol{x}^+)\rangle] &= -\mathbb{E}[\langle \boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x}), \nabla\boldsymbol{y}^\star(\boldsymbol{x})(\boldsymbol{x}^+ - \boldsymbol{x})\rangle] \\ &\quad - \mathbb{E}[\langle \boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x}), \boldsymbol{y}^\star(\boldsymbol{x}^+) - \boldsymbol{y}^\star(\boldsymbol{x}) - \nabla\boldsymbol{y}^\star(\boldsymbol{x})(\boldsymbol{x}^+ - \boldsymbol{x})\rangle].\end{aligned} \tag{30}$$

Recall that $\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+) = \mathbb{E}\left[\boldsymbol{h}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+)|\mathcal{F}_{i,\ell-1}\right]$. Thus,

$$\begin{aligned}-\mathbb{E}[\langle \boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x}), \nabla\boldsymbol{y}^\star(\boldsymbol{x})(\boldsymbol{x}^+ - \boldsymbol{x})\rangle] &= -\mathbb{E}\left[\langle \boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x}), \nabla\boldsymbol{y}^\star(\boldsymbol{x})\sum_{i=1}^m \alpha_i \sum_{\ell=0}^{\tau_i-1} \bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+)\rangle\right] \\ &\leq \mathbb{E}\left[\|\boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x})\|\|\frac{1}{m}\nabla\boldsymbol{y}^\star(\boldsymbol{x})\sum_{i=1}^m \alpha_i \sum_{\ell=0}^{\tau_i-1} \bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+)\|\right] \\ &\leq L_y\mathbb{E}\left[\|\boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x})\|\|\frac{1}{m}\sum_{i=1}^m \alpha_i \sum_{\ell=0}^{\tau_i-1} \bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+)\|\right] \\ &\leq 2\gamma\mathbb{E}\left[\left\|\boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x})\right\|^2\right] + \frac{L_y^2\alpha^2}{8\gamma}\mathbb{E}\left[\|\frac{1}{m}\sum_{i=1}^m \frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1} \bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+)\|^2\right],\end{aligned} \tag{31}$$

where the first equality uses the fact that $\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell}\boldsymbol{y}^+) = \mathbb{E}\left[\boldsymbol{h}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+)|\mathcal{F}_{i,\ell-1}\right]$; the second inequality follows from Lemma B.1; and the last inequality follows from the Young's inequality such that $ab \leq \gamma a^2 + \frac{b^2}{4\gamma}$.

Further, using Lemma B.1, we have

$$\begin{aligned}&-\mathbb{E}[\langle \boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x}), \boldsymbol{y}^\star(\boldsymbol{x}^+) - \boldsymbol{y}^\star(\boldsymbol{x}) - \nabla\boldsymbol{y}^\star(\boldsymbol{x})(\boldsymbol{x}^+ - \boldsymbol{x})\rangle] \\ &\leq \mathbb{E}\left[\left\|\boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x})\right\|\left\|\boldsymbol{y}^\star(\boldsymbol{x}^+) - \boldsymbol{y}^\star(\boldsymbol{x}) - \nabla\boldsymbol{y}^\star(\boldsymbol{x})(\boldsymbol{x}^+ - \boldsymbol{x})\right\|\right] \\ &\leq \frac{L_{yx}}{2}\mathbb{E}\left[\left\|\boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x})\right\|\left\|\boldsymbol{x}^+ - \boldsymbol{x}\right\|^2\right],\end{aligned} \tag{32}$$

where the inequality follows from Lemma B.1.

From Algorithm 2, we have

$$\boldsymbol{x}_{i,0} = \boldsymbol{x}, \forall i \in \mathcal{S}, \quad \text{and} \quad \boldsymbol{x}^+ = \boldsymbol{x} - \frac{1}{m}\sum_{i=1}^{m}\frac{\alpha}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\boldsymbol{h}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+). \tag{33}$$

Let

$$A := \mathbb{E}\left[\mathbb{E}\left[\left\|\boldsymbol{x}^+ - \boldsymbol{x}\right\|^2 \mid \mathcal{F}_{i,0}\right] \mid \ldots \mid \mathcal{F}_{i,\tau_i-1}\right]. \tag{34}$$

From Lemma B.1, we get

$$A \leq \alpha^2 \tilde{D}_f^2. \tag{35}$$

Note that for any $\eta > 0$, we have $1 \leq \frac{\eta}{2} + \frac{1}{2\eta}$. Combining this inequality with (32), and using (35) give

$$\begin{aligned}
&-\mathbb{E}[\langle \boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x}), \boldsymbol{y}^\star(\boldsymbol{x}^+) - \boldsymbol{y}^\star(\boldsymbol{x}) - \nabla\boldsymbol{y}^\star(\boldsymbol{x})(\boldsymbol{x}^+ - \boldsymbol{x})\rangle] \\
&\leq \frac{\eta L_{yx}}{4}\mathbb{E}\left[\left\|\boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x})\right\|^2 A\right] + \frac{L_{yx}}{4\eta}\mathbb{E}\left[A\right] \\
&\leq \frac{\eta L_{yx}\tilde{D}_f^2\alpha^2}{4}\mathbb{E}\left[\left\|\boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x})\right\|^2\right] + \frac{L_{yx}}{4\eta}\mathbb{E}\left[A\right] \\
&\leq \frac{\eta L_{yx}\tilde{D}_f^2\alpha^2}{4}\mathbb{E}\left[\left\|\boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x})\right\|^2\right] \\
&+ \frac{L_{yx}\alpha^2}{4\eta}\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+)\right\|^2\right] + \frac{L_{yx}\alpha^2}{4\eta}\tilde{\sigma}_f,
\end{aligned} \tag{36}$$

where the last inequality uses Lemma F.4.

Let $\gamma = M_f L_y \alpha$. Plugging (36) and (31) into (30), we have

$$\begin{aligned}
\mathbb{E}[\langle \boldsymbol{y}^+ - \boldsymbol{y}^*(\boldsymbol{x}), \boldsymbol{y}^*(\boldsymbol{x}) - \boldsymbol{y}^*(\boldsymbol{x}^+)\rangle] &\leq \left(2\gamma + \frac{\eta L_{yx}\tilde{D}_f^2}{4}\alpha^2\right)\mathbb{E}\left[\left\|\boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x})\right\|^2\right] \\
&+ \left(\frac{L_y^2\alpha^2}{8\gamma} + \frac{L_{yx}\alpha^2}{4\eta}\right)\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+)\right\|^2\right] \\
&= \left(2M_f L_y\alpha + \frac{\eta L_{yx}\tilde{D}_f^2}{4}\alpha^2\right)\mathbb{E}\left[\left\|\boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x})\right\|^2\right] \\
&+ \left(\frac{L_y\alpha}{8M_f} + \frac{L_{yx}\alpha^2}{4\eta}\right)\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+)\right\|^2\right].
\end{aligned} \tag{37}$$

Substituting (37) and (29) into (27) completes the proof. □

B.3. Drifting Errors of FEDOUT

The following lemma provides a bound on the *drift* of each $\boldsymbol{x}_{i,\ell}$ from $\boldsymbol{x}$ for general stochastic nonconvex problems. It should be mentioned that similar drifting bounds for *single-level* problems are provided under either strong convexity (Mitra et al., 2021) and/or bounded dissimilarity assumptions (Wang et al., 2020; Reddi et al., 2020; Li et al., 2020b).

Lemma B.4 (Drifting Error of FEDOUT)**.** *Suppose Assumptions A-B hold. Further, assume* $\tau_i \geq 1$ *and* $\alpha_i \leq 1/(8M_f\tau_i), \forall i \in \mathcal{S}$*. Then,* $\forall \ell \in \{0, \ldots, \tau_i - 1\}$, FEDOUT *guarantees:*

$$\mathbb{E}\left[\left\|\boldsymbol{x}_{i,\ell} - \boldsymbol{x}\right\|^2\right] \leq 30\tau_i^2\alpha_i^2\mathbb{E}\left[\left\|\nabla f(\boldsymbol{x})\right\|^2\right] + 15\tau_i^2\alpha_i^2(3\tilde{\sigma}_f^2 + 6b^2) + 30\tau_i^2\alpha_i^2\|\boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x})\|^2. \tag{38}$$

Proof. The result trivially holds for $\tau_i = 1$. Let $\tau_i > 1$ and define

$$\begin{aligned}
\boldsymbol{v}_{i,\ell} &:= \bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+) - \bar{\nabla} f_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+) - \bar{\boldsymbol{h}}_i(\boldsymbol{x}, \boldsymbol{y}^+) \\
&\quad + \bar{\nabla} f_i(\boldsymbol{x}, \boldsymbol{y}^+) + \bar{\boldsymbol{h}}(\boldsymbol{x}, \boldsymbol{y}^+) - \bar{\nabla} f(\boldsymbol{x}, \boldsymbol{y}^+), \\
\boldsymbol{w}_{i,\ell} &:= \boldsymbol{h}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+) - \bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+) + \bar{\boldsymbol{h}}_i(\boldsymbol{x}, \boldsymbol{y}^+) \\
&\quad - \boldsymbol{h}_i(\boldsymbol{x}, \boldsymbol{y}^+) + \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}^+) - \bar{\boldsymbol{h}}(\boldsymbol{x}, \boldsymbol{y}^+), \\
\boldsymbol{z}_{i,\ell} &:= \bar{\nabla} f_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+) - \bar{\nabla} f_i(\boldsymbol{x}, \boldsymbol{y}^+) + \bar{\nabla} f(\boldsymbol{x}, \boldsymbol{y}^+) - \nabla f(\boldsymbol{x}) + \nabla f(\boldsymbol{x}).
\end{aligned} \tag{39}$$

From Algorithm 2, for each $i \in \mathcal{S}$, and $\forall \ell \in \{0, \ldots, \tau_i - 1\}$, we have

$$\boldsymbol{x}_{i,\ell+1} - \boldsymbol{x} = (\boldsymbol{x}_{i,\ell} - \boldsymbol{x}) - \alpha_i(\boldsymbol{v}_{i,\ell} + \boldsymbol{w}_{i,\ell} + \boldsymbol{z}_{i,\ell}), \tag{40}$$

which implies that

$$\begin{aligned}
\mathbb{E}\big[\|\boldsymbol{x}_{i,\ell+1} - \boldsymbol{x}\|^2\big] &= \mathbb{E}\big[\|(\boldsymbol{x}_{i,\ell} - \boldsymbol{x} - \alpha_i(\boldsymbol{v}_{i,\ell} + \boldsymbol{z}_{i,\ell})\|^2\big] + \alpha_i^2 \mathbb{E}\big[\|\boldsymbol{w}_{i,\ell}\|^2\big] \\
&\quad - 2\mathbb{E}\Big[\mathbb{E}\Big[\langle \boldsymbol{x}_{i,\ell} - \boldsymbol{x} - \alpha_i(\boldsymbol{v}_{i,\ell} + \boldsymbol{z}_{i,\ell}), \alpha_i \boldsymbol{w}_{i,\ell}\rangle | \mathcal{F}_{i,\ell-1}\Big]\Big] \\
&= \mathbb{E}\big[\|\boldsymbol{x}_{i,\ell} - \boldsymbol{x} - \alpha_i(\boldsymbol{v}_{i,\ell} + \boldsymbol{z}_{i,\ell})\|^2\big] + \alpha_i^2 \mathbb{E}\big[\|\boldsymbol{w}_{i,\ell}\|^2\big].
\end{aligned} \tag{41}$$

Here, the last equality uses Lemma F.3 since $\mathbb{E}[\boldsymbol{w}_{i,\ell} | \mathcal{F}_{i,\ell-1}] = 0$, by definition.

Observe that

$$\begin{aligned}
\mathbb{E}\Big[\|\boldsymbol{v}_{i,\ell}\|^2\Big] &\le 3\mathbb{E}\Big[\|\bar{\nabla} f_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+) - \bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+)\|^2 \\
&\quad + \|\bar{\boldsymbol{h}}_i(\boldsymbol{x}, \boldsymbol{y}^+) - \bar{\nabla} f_i(\boldsymbol{x}, \boldsymbol{y}^+)\|^2 + \|\bar{\nabla} f(\boldsymbol{x}, \boldsymbol{y}^+) - \bar{\boldsymbol{h}}(\boldsymbol{x}, \boldsymbol{y}^+)\|^2\Big] \\
&\le 9b^2,
\end{aligned} \tag{42a}$$

where the first inequality follows from (39) and Lemma F.1 and the last inequality uses Lemma F.4.

Similarly, we obtain

$$\begin{aligned}
\mathbb{E}\Big[\|\boldsymbol{w}_{i,\ell}\|^2\Big] &\le 3\mathbb{E}\Big[\|\boldsymbol{h}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+) - \bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+)\|^2 \\
&\quad + \|\bar{\boldsymbol{h}}_i(\boldsymbol{x}) - \boldsymbol{h}_i(\boldsymbol{x}, \boldsymbol{y}^+)\|^2 + \|\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}^+) - \bar{\boldsymbol{h}}(\boldsymbol{x}, \boldsymbol{y}^+)\|^2\Big] \\
&\le 9\tilde{\sigma}_f^2.
\end{aligned} \tag{42b}$$

The first term in the RHS of (41) can be bounded as follows:

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{x}_{i,\ell+1} - \boldsymbol{x}\|^2\right] &\le (1 + \frac{1}{2\tau_i - 1})\mathbb{E}\Big[\|\boldsymbol{x}_{i,\ell} - \boldsymbol{x}\|^2\Big] + 2\tau_i \mathbb{E}\Big[\|\alpha_i(\boldsymbol{v}_{i,\ell} + \boldsymbol{z}_{i,\ell})\|^2\Big] \\
&\le (1 + \frac{1}{2\tau_i - 1})\mathbb{E}\Big[\|\boldsymbol{x}_{i,\ell} - \boldsymbol{x}\|^2\Big] + 2\tau_i \alpha_i^2 \mathbb{E}\Big[\|\boldsymbol{z}_{i,\ell}\|^2\Big] + 18\tau_i \alpha_i^2 b^2 \\
&\le (1 + \frac{1}{2\tau_i - 1} + 6\tau_i \alpha_i^2 M_f^2)\mathbb{E}\Big[\|\boldsymbol{x}_{i,\ell} - \boldsymbol{x}\|^2\Big] \\
&\quad + 6\tau_i \alpha_i^2 \mathbb{E}\Big[\|\nabla f(\boldsymbol{x})\|^2\Big] + 18\tau_i \alpha_i^2 b^2 + 6\tau_i \alpha_i^2 \|\boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x})\|^2.
\end{aligned} \tag{43}$$

Here, the first inequality follows from Lemma F.2; the second and last inequalities follow from (39), (79), and Lemma B.1.

Substituting (42) and (43) into (41) gives

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{x}_{i,\ell+1} - \boldsymbol{x}\|^2\right] &\le (1 + \frac{1}{2\tau_i - 1} + 6\tau_i \alpha_i^2 M_f^2)\mathbb{E}\Big[\|\boldsymbol{x}_{i,\ell} - \boldsymbol{x}\|^2\Big] \\
&\quad + 6\tau_i \alpha_i^2 \mathbb{E}\Big[\|\nabla f(\boldsymbol{x})\|^2\Big] + 3\tau_i \alpha_i^2 (3\tilde{\sigma}_f^2 + 6b^2) + 6\tau_i \alpha_i^2 \|\boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x})\|^2 \\
&\le (1 + \frac{1}{\tau_i - 1})\mathbb{E}\Big[\|\boldsymbol{x}_{i,\ell} - \boldsymbol{x}\|^2\Big] \\
&\quad + 6\tau_i \alpha_i^2 \mathbb{E}\Big[\|\nabla f(\boldsymbol{x})\|^2\Big] + 3\tau_i \alpha_i^2 (3\tilde{\sigma}_f^2 + 6b^2) + 6\tau_i \alpha_i^2 \|\boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x})\|^2
\end{aligned} \tag{44}$$

where the last inequality follows by noting $\alpha_i \leq 1/(8M_f\tau_i)$.

Let

$$\omega_l = \sum_{j=0}^{\ell-1} \omega_i^j \;\text{ where }\; \omega_i := 1 + \frac{1}{\tau_i - 1}.$$

Since $\tau_i > 1$, we have

$$\omega_l = \frac{\omega_i^\ell - 1}{\omega_i - 1} \leq 5(\tau_i - 1).$$

Now, iterating equation (44) and using $\boldsymbol{x}_{i,0} = \boldsymbol{x}$, we obtain:

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{x}_{i,\ell} - \boldsymbol{x}\|^2\right] &\leq \left(6\tau_i\alpha_i^2\mathbb{E}\left[\|\nabla f(\boldsymbol{x})\|^2\right] + 3\tau_i\alpha_i^2(3\tilde{\sigma}_f^2 + 6b^2) + +6\tau_i\alpha_i^2\|\boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x})\|^2\right)\omega_l \\ &\leq 30\tau_i^2\alpha_i^2\mathbb{E}\left[\|\nabla f(\boldsymbol{x})\|^2\right] + 15\tau_i^2\alpha_i^2\left(3\tilde{\sigma}_f^2 + 6b^2\right) + 30\tau_i^2\alpha_i^2\|\boldsymbol{y}^+ - \boldsymbol{y}^\star(\boldsymbol{x})\|^2.\end{aligned} \tag{45}$$

This completes the proof. □

From Lemma B.4, we have $\|\boldsymbol{x}_{i,\ell} - \boldsymbol{x}\| \leq \mathcal{O}(\tau_i\alpha_i)$ which shows that the bound on the *client-drift* scales linearly with τ_i in general nested FL. Hence, to cancel out such a drift, one may choose $\alpha_i = \mathcal{O}(1/\tau_i)$ for all $i \in \mathcal{S}$.

B.4. Proof of Theorem 3.1

Here, we provide the proof of our main result which can be adapted to general nested problems (bilevel, min-max, composite). Additionally, when there is no inner problem, the setting reduces to single-level optimization and the result leads to a new convergence guarantee for federated non-convex variance reduction methods (compared to the recent strongly-convex results of FEDLIN (Mitra et al., 2021)).

Proof. Following (Chen et al., 2021a), we define the following Lyapunov function

$$\mathbb{W}^k := f(\boldsymbol{x}^k) + \frac{M_f}{L_y}\|\boldsymbol{y}^k - \boldsymbol{y}^\star(\boldsymbol{x}^k)\|^2. \tag{46}$$

In the following, we bound the difference between two Lyapunov functions

$$\mathbb{W}^{k+1} - \mathbb{W}^k = f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^k) \; + \; \frac{M_f}{L_y}(\|\boldsymbol{y}^{k+1} - \boldsymbol{y}^*(\boldsymbol{x}^{k+1})\|^2 - \|\boldsymbol{y}^k - \boldsymbol{y}^\star(\boldsymbol{x}^k)\|^2).$$

From Lemmas B.2 and B.3, we have

$$\begin{aligned}\mathbb{E}[\mathbb{W}^{k+1}] - \mathbb{E}[\mathbb{W}^k] \leq & \underbrace{-\frac{\alpha_k}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] + \frac{3M_f^2\alpha_k}{2m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\mathbb{E}\left[\|\boldsymbol{x}_{i,\ell}^k - \boldsymbol{x}^k\|^2\right]}_{\mathcal{A}} \\ & - \left(\frac{\alpha_k}{2} - \frac{L_f\alpha_k^2}{2} - \frac{M_f}{L_{\boldsymbol{y}}}a_1(\alpha_k)\right)\mathbb{E}\left[\|\frac{1}{m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell}^k, \boldsymbol{y}^{k+1})\|^2\right] \\ & \underbrace{+\frac{M_f(M_fL_y\alpha_k + a_2(\alpha_k))}{L_y}\mathbb{E}[\|\boldsymbol{y}^{k+1} - \boldsymbol{y}^*(\boldsymbol{x}^k)\|^2] - \frac{M_f}{L_y}\mathbb{E}[\|\boldsymbol{y}^k - \boldsymbol{y}^*(\boldsymbol{x}^k)\|^2]}_{\mathcal{B}} \\ & + \left(\frac{L_f\alpha_k^2}{2} + a_3(\alpha_k)\right)\tilde{\sigma}_f^2 + \frac{3\alpha_k}{2}b_k^2.\end{aligned} \tag{47}$$

where $a_1(\alpha) - a_3(\alpha)$ are defined in (26).

Using Lemma B.4, we obtain

$$\begin{aligned}\mathcal{A} &\leq -\frac{\alpha_k}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] + \frac{3M_f^2\alpha_k^3}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] + \frac{45M_f^2\alpha_k^3}{2}\left(3\tilde{\sigma}_f^2 + 8b_k^2\right) + 30\alpha_k^3\|\boldsymbol{y}^{k+1} - \boldsymbol{y}^\star(\boldsymbol{x})\|^2 \\ &\leq -\frac{\alpha_k}{4}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] + \frac{45M_f^2\alpha_k^3}{2}\left(3\tilde{\sigma}_f^2 + 6b^2\right) + 30\alpha_k^3\|\boldsymbol{y}^{k+1} - \boldsymbol{y}^\star(\boldsymbol{x})\|^2 \\ &\leq -\frac{\alpha_k}{4}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] + 3\alpha_k^2\left(3\tilde{\sigma}_f^2 + 6b_k^2\right) + 30\alpha_k^3\|\boldsymbol{y}^{k+1} - \boldsymbol{y}^\star(\boldsymbol{x})\|^2.\end{aligned} \tag{48}$$

Here, we use our assumption that

$$\alpha_k \leq \frac{1}{8}M_f. \tag{49}$$

Further, using Lemma B.3, we have

$$\begin{aligned}\mathcal{B} \leq \frac{M_f}{L_y}&\left((M_fL_y\alpha_k + a_2(\alpha_k))\left(1 - \frac{\beta_k\mu_g}{2}\right)^T - 1\right)\mathbb{E}[\|\boldsymbol{y}^k - \boldsymbol{y}^*(\boldsymbol{x}^k)\|^2] \\ &+ \frac{25M_fa_2(\alpha_k)}{L_y}T\beta_k^2\sigma_{g,1}^2.\end{aligned} \tag{50}$$

Substituting (50) and (48) into (47) gives

$$\begin{aligned}\mathbb{E}[\mathbb{W}^{k+1}] - \mathbb{E}[\mathbb{W}^k] \leq &-\frac{\alpha_k}{4}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] \\ &- \left(\frac{\alpha_k}{2} - \frac{L_f\alpha_k^2}{2} - \frac{M_f}{L_{\boldsymbol{y}}}a_1(\alpha_k)\right)\mathbb{E}\left[\|\frac{1}{m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell}^k, \boldsymbol{y}^{k+1})\|^2\right] \\ &+ \frac{M_f}{L_y}\left((1 + \frac{30L_y}{M_f}\alpha_k^3 + M_fL_y\alpha_k + a_2(\alpha_k))\left(1 - \frac{\beta_k\mu_g}{2}\right)^T - 1\right)\mathbb{E}[\|\boldsymbol{y}^k - \boldsymbol{y}^*(\boldsymbol{x}^k)\|^2] \\ &+ \left(\frac{L_f\alpha_k^2}{2} + a_3(\alpha_k) + 9\alpha_k^2\right)\tilde{\sigma}_f^2 \\ &+ \left(\frac{3\alpha_k}{2} + 24\alpha_k^2\right)b_k^2 \\ &+ \frac{25M_f}{L_y}\left(\frac{30L_y}{M_f}\alpha_k^3 + M_fL_y\alpha_k + a_2(\alpha_k)\right)T\beta_k^2\sigma_{g,1}^2.\end{aligned} \tag{51}$$

To guarantee the descent of $\mathbb{W}^k$, the following constraints need to be satisfied

$$\begin{aligned}&\frac{\alpha_k}{2} - \frac{L_f\alpha_k^2}{2} - \frac{M_f}{L_{\boldsymbol{y}}}a_1(\alpha_k) \geq 0, \\ \implies &\frac{\alpha_k}{2} - \frac{L_f\alpha_k^2}{2} - \frac{M_f}{L_{\boldsymbol{y}}}\left(L_y\alpha_k^2 + \frac{L_y\alpha_k}{4M_f} + \frac{L_{yx}\alpha_k^2}{2\eta}\right) \geq 0, \\ \implies &\alpha_k \leq \frac{1}{2L_f + 4M_fL_y + \frac{M_fL_{yx}}{L_y\eta}}\end{aligned} \tag{52}$$

where the second line uses (26).

Similarly, for any $\alpha_k > 0$, we have

$$\begin{aligned}&\frac{M_f}{L_y}\left((1 + \frac{30L_y}{M_f}\alpha_k^3 + M_fL_y\alpha_k + a_2(\alpha_k))\left(1 - \frac{\beta_k\mu_g}{2}\right)^T - 1\right) \leq 0, \\ \implies &\frac{M_f}{L_y}\left(\left(1 + \frac{30L_y}{M_f}\alpha_k^3 + 5M_fL_y\alpha_k + \frac{\eta L_{yx}\tilde{D}_f^2\alpha_k^2}{2}\right)\left(1 - \frac{\beta_k\mu_g}{2}\right)^T - 1\right) \leq 0, \\ \implies &\beta_k \geq \frac{5M_fL_y + \frac{\eta L_{yx}\tilde{D}_f^2}{4}\alpha_k + \frac{30L_y}{M_f}\alpha_k^2}{T\mu_g}\alpha_k,\end{aligned} \tag{53}$$

where the second line uses (26).

From (49), (52) and (53), we select

$$\begin{aligned}\bar{\alpha}_1 &:= \frac{1}{2L_f + 4M_f L_y + \frac{M_f L_{yx}}{L_y \eta}},\\ \bar{\alpha}_2 &:= \frac{T\rho_g}{8(\mu_g + \ell_{g,1})(40M_f L_y + 2\eta L_{yx}\tilde{C}_f^2 \bar{\alpha}_1)}, \quad \bar{\alpha}_3 := \frac{1}{8M_f},\\ \bar{\beta} &:= \frac{1}{4\rho_g}\left(20M_f L_y + \eta L_{yx}\tilde{D}_f^2\bar{\alpha}_1 + \frac{120L_y}{M_f}\bar{\alpha}_1^2\right).\end{aligned} \tag{54}$$

where $\rho_g := \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}$.

To satisfy (52) and (53), we choose

$$\alpha_k = \min\{\bar{\alpha}_1, \bar{\alpha}_2, \bar{\alpha}_3, \frac{\bar{\alpha}}{\sqrt{K}}\}, \qquad \beta_k = \frac{\bar{\beta}\alpha_k}{T}. \tag{55}$$

With the above choice of stepsizes, (51) can be simplified as

$$\begin{aligned}\mathbb{E}[\mathbb{W}^{k+1}] - \mathbb{E}[\mathbb{W}^k] &\le -\frac{\alpha_k}{4}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2]\\ &+ \left(\frac{L_f\alpha_k^2}{2} + a_3(\alpha_k) + 9\alpha_k^2\right)\tilde{\sigma}_f^2\\ &+ \left(\frac{3\alpha_k}{2} + 24\alpha_k^2\right)b_k^2\\ &+ \frac{25M_f}{L_y}\left(M_f L_y \alpha_k + a_2(\alpha_k)\right)\sigma_{g,1}^2\\ &\le -\frac{\alpha_k}{4}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] + c_1\alpha_k^2\sigma_{g,1}^2 + \left(\frac{3\alpha_k}{2} + 24\alpha_k^2\right)b_k^2 + c_2\alpha_k^2\tilde{\sigma}_f^2,\end{aligned} \tag{56}$$

where the constants c_1 and c_2 are defined as

$$\begin{aligned}c_1 &= \frac{25M_f}{L_y}\left(1 + 2M_f L_y\bar{\alpha}_1 + \frac{\eta L_{yx}\tilde{D}_f^2}{4}\bar{\alpha}_1^2 + \frac{30L_y}{M_f}\bar{\alpha}_1^3\right)\bar{\beta}^2\frac{1}{T},\\ c_2 &= \frac{L_f}{2} + M_f L_y + \frac{L_{yx}M_f}{4\eta L_y} + 9.\end{aligned} \tag{57}$$

Then telescoping gives

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] &\le \frac{\mathbb{W}^0 + \sum_{k=0}^{K-1}\alpha_k b_k^2 + c_1\alpha_k^2\sigma_{g,1}^2 + c_2 T\beta_k^2\tilde{\sigma}_f^2}{\frac{1}{2}\sum_{k=0}^{K-1}\alpha_k}\\ &\le \frac{2\mathbb{W}^0}{K\min\{\bar{\alpha}_1, \bar{\alpha}_2, \bar{\alpha}_3\}} + \frac{2\mathbb{W}^0}{\bar{\alpha}\sqrt{K}}\\ &+ (\frac{3}{2M_f} + \frac{3}{M_f^2})b_k^2 + \frac{2c_1\bar{\alpha}}{\sqrt{K}}\sigma_{g,1}^2 + \frac{2c_2\bar{\alpha}}{\sqrt{K}}\tilde{\sigma}_f^2.\end{aligned} \tag{58}$$

This completes the proof. □

B.5. Proof of Corollary 3.1

Proof. Let $\eta = \frac{M_f}{L_y} = \mathcal{O}(\kappa_g)$ in (57). We can have

$$\bar{\alpha}_1 = \bar{\alpha}_3 = \mathcal{O}(\kappa_g^{-3}), \quad \bar{\alpha}_2 = \mathcal{O}(T\kappa_g^{-3}), \quad c_1 = \mathcal{O}(\kappa_g^9/T), \quad c_2 = \mathcal{O}(\kappa_g^3) \tag{59}$$

Further, $N = \mathcal{O}(\kappa_g \log K)$ gives $b_k = \frac{1}{K^{1/4}}$. Now, if we select $\bar{\alpha} = \mathcal{O}(\kappa_g^{-2.5})$ and $T = \mathcal{O}(\kappa_g^4)$, then

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] = \mathcal{O}\left(\frac{\kappa_g^3}{K} + \frac{\kappa_g^{2.5}}{\sqrt{K}}\right).$$

To achieve ε-optimal solution, we need $K = \mathcal{O}(\kappa_g^5 \varepsilon^{-2})$. □

C. Proof for Federated Min-Max Optimization

Note that the inner function for the min-max problem is $g_i(\boldsymbol{x}, \boldsymbol{y}; \xi) = -f_i(\boldsymbol{x}, \boldsymbol{y}; \xi)$. Then, the problem (1) has the following form

$$\begin{array}{ll} \min\limits_{\boldsymbol{x}\in\mathbb{R}^{d_1}} & f(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^m f_i\left(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})\right) \\ \text{subj. to} & \boldsymbol{y}^*(\boldsymbol{x}) = \underset{\boldsymbol{y}\in\mathbb{R}^{d_2}}{\operatorname{argmin}} \ -\frac{1}{m}\sum_{i=1}^m f_i\left(\boldsymbol{x}, \boldsymbol{y}\right). \end{array} \tag{60a}$$

Here,

$$f_i(\boldsymbol{x}, \boldsymbol{y}) := \mathbb{E}_{\xi\sim\mathcal{C}_i}[f_i(\boldsymbol{x}, \boldsymbol{y}; \xi)]$$

is the loss functions of the i^{th} client.

One can notice that the hypergradient of (60) has the following form

$$\nabla f_i(\boldsymbol{x}) := \nabla_{\boldsymbol{x}} f_i\big(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})\big) + \nabla_{\boldsymbol{x}}\boldsymbol{y}^*(\boldsymbol{x})^\top \nabla_{\boldsymbol{y}} f_i\big(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})\big) = \nabla_{\boldsymbol{x}} f_i\big(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})\big), \tag{61}$$

where the second equality follows from the optimality condition of the inner problem, i.e., $\nabla_y f(x, y^*(x)) = 0$. For each $i \in \mathcal{S}$, we can approximate $\nabla f_i(\boldsymbol{x})$ on a vector $\boldsymbol{y}$ in place of $\boldsymbol{y}^*(\boldsymbol{x})$, denoted as $\overline{\nabla} f_i(\boldsymbol{x}, \boldsymbol{y}) := \nabla_{\boldsymbol{x}} f_i\big(\boldsymbol{x}, \boldsymbol{y}\big)$. Note that in the min-max case $\boldsymbol{h}_i$ is an unbiased estimator of $\bar{\nabla} f_i(\boldsymbol{x}, \boldsymbol{y})$. Thus, $b = 0$.

Therefore, the alternating stochastic gradients for this special case are given by

$$\begin{aligned} \boldsymbol{q}_{i,\ell} &= -\nabla_y f_i(\boldsymbol{x}, \boldsymbol{y}_{i,\ell}; \xi_{i,\ell}) + \nabla_y f_i(\boldsymbol{x}, \boldsymbol{y}; \xi_{i,\ell}) + \boldsymbol{q}(\boldsymbol{x}, \boldsymbol{y}), \\ \boldsymbol{h}_{i,\ell} &= -\nabla_x f_i(\boldsymbol{x}_{i,\ell}, \boldsymbol{y}^+; \xi_{i,\ell}) + \nabla_x f_i(\boldsymbol{x}, \boldsymbol{y}^+; \xi_{i,\ell}) + \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}^+). \end{aligned} \tag{62}$$

C.1. Supporting Lemmas

Let $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{d_1+d_2}$. We make the following assumptions that are counterparts of Assumptions A and B.

Assumption E. *For all $i \in [m]$:*

(C1) *$f_i(\boldsymbol{z}), \nabla f_i(\boldsymbol{z}), \nabla^2 f_i(\boldsymbol{z})$ are respectively $\ell_{f,0}$, $\ell_{f,1}, \ell_{f,2}$-Lipschitz continuous; and*

(C2) *$f_i(\boldsymbol{x}, \boldsymbol{y})$ is μ_f-strongly convex in $\boldsymbol{y}$ for any fixed $\boldsymbol{x} \in \mathbb{R}^{d_1}$.*

We use $\kappa = \ell_{f,1}/\mu_f$ to denote the condition number of the inner objective with respect to $\boldsymbol{y}$.

Assumption F. *For all $i \in [m]$:*

(D1) *$\nabla f_i(\boldsymbol{z}; \xi)$ is unbiased estimators of $\nabla f_i(\boldsymbol{z})$; and*

(D2) *Its variance is bounded, i.e., $\mathbb{E}_\xi[\|\nabla f_i(\boldsymbol{z}; \xi) - \nabla f_i(\boldsymbol{z})\|^2] \le \sigma_f^2$, for some σ_f^2.*

Lemma C.1. *Under Assumption E and F, we have*

$$\|\nabla f(\boldsymbol{x}_1) - \nabla f(\boldsymbol{x}_2)\| \le L_f\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|, \tag{63a}$$
$$\|\boldsymbol{y}^*(\boldsymbol{x}_1) - \boldsymbol{y}^*(\boldsymbol{x}_2)\| \le L_y\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|, \tag{63b}$$
$$\|\nabla\boldsymbol{y}^*(\boldsymbol{x}_1) - \nabla\boldsymbol{y}^*(\boldsymbol{x}_2)\| \le L_{yx}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|, \tag{63c}$$
$$\tag{63d}$$

and for all $i \in \mathcal{S}$, we have

$$\|\bar{\nabla} f_i(\boldsymbol{x},\boldsymbol{y}) - \nabla f_i(\boldsymbol{x},\boldsymbol{y}^*(\boldsymbol{x}))\| \le \bar{L}_f \|\boldsymbol{y}^*(\boldsymbol{x}) - \boldsymbol{y}\|, \text{ for all } i \in \{1,\dots,m\}, \tag{63e}$$

$$\mathbb{E}\left[\|\bar{\boldsymbol{h}}_i(\boldsymbol{x},\boldsymbol{y}) - \boldsymbol{h}_i(\boldsymbol{x},\boldsymbol{y})\|^2\right] \le \tilde{\sigma}_f^2, \tag{63f}$$

$$\mathbb{E}\left[\|\boldsymbol{h}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+)\|^2 | \mathcal{F}_{i,\ell-1}\right] \le \tilde{D}_f. \tag{63g}$$

Here,

$$L_{yx} = \frac{\ell_{f,2} + \ell_{f,2}L_y}{\mu_f} + \frac{\ell_{f,1}(\ell_{f,2} + \ell_{f,2}L_y)}{\mu_f^2} = \mathcal{O}(\kappa_f^3),$$
$$M_f = \ell_{f,1} = \mathcal{O}(1), \quad L_f = (\ell_{f,1} + \frac{\ell_{f,1}^2}{\mu_f}) = \mathcal{O}(\kappa_f),$$
$$L_y = \frac{\ell_{f,1}}{\mu_f} = \mathcal{O}(\kappa), \quad \tilde{\sigma}_f^2 = \sigma_f^2, \quad \tilde{D}_f^2 = \ell_{l,0}^2 + \sigma_f^2,$$

where $\ell_{f,0}$, $\ell_{f,1}, \ell_{f,2}$, and μ_f are given in Assumption E.

C.2. Proof of Corollary 3.2

Proof. Let

$$\begin{aligned}
\bar{\alpha}_1 &:= \frac{1}{2L_f + 4M_f L_y + \frac{M_f L_{yx}}{L_y}}, \\
\bar{\alpha}_2 &:= \frac{T\rho_g}{8(\mu_g + \ell_{g,1})(40M_f L_y + 2L_{yx}\tilde{C}_f^2\bar{\alpha}_1)}, \quad \bar{\alpha}_3 := \frac{1}{8M_f}, \\
\bar{\beta} &:= \frac{1}{4\rho_g}\left(20M_f L_y + L_{yx}\tilde{D}_f^2\bar{\alpha}_1 + \frac{120L_y}{M_f}\bar{\alpha}_1^2\right).
\end{aligned} \tag{64}$$

where $\rho_g := \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}$.

We select

$$\alpha_k = \min\{\bar{\alpha}_1, \bar{\alpha}_2, \bar{\alpha}_3, \frac{\bar{\alpha}}{\sqrt{K}}\}, \qquad \beta_k = \frac{\bar{\beta}\alpha_k}{T}. \tag{65}$$

Using the above choice of stepsizes, (58) reduces to

$$\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] \le \frac{2\mathbb{W}^0}{K\min\{\bar{\alpha}_1,\bar{\alpha}_2,\bar{\alpha}_3\}} + \frac{2\mathbb{W}^0}{\bar{\alpha}\sqrt{K}} \\
+ \frac{2(c_1+c_2)\bar{\alpha}}{\sqrt{K}}\sigma_f^2.
\end{aligned} \tag{66}$$

where

$$c_1 = \frac{25M_f}{L_y}\left(1 + 2M_f L_y\bar{\alpha}_1 + \frac{L_{yx}\tilde{D}_f^2}{4}\bar{\alpha}_1^2 + \frac{30L_y}{M_f}\bar{\alpha}_1^3\right)\bar{\beta}^2\frac{1}{T} = \mathcal{O}(\frac{\kappa^3}{T})$$
$$c_2 = \frac{L_f}{2} + M_f L_y + \frac{L_{yx}M_f}{4L_y} + 9 = \mathcal{O}(\kappa^2).$$

Note that $\bar{\alpha}_1 = \bar{\alpha}_3 = \mathcal{O}(\kappa_f^{-2})$, $\bar{\alpha}_2 = \mathcal{O}(T\kappa_f^{-2})$. Let $\bar{\alpha} = \mathcal{O}(\kappa_f^{-1})$, $T = \mathcal{O}(\kappa_f)$, we have

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] = \mathcal{O}\left(\frac{\kappa_f^2}{K} + \frac{\kappa_f}{\sqrt{K}}\right). \tag{67}$$

To achieve ε-accuracy, we need $K = \mathcal{O}(\kappa_f^2\varepsilon^{-2})$. □

D. Proof for Federated Single-Level Optimization

Next we re-derive Lemmas Lemmas B.2 and B.4 for single-level nonconvex FL under some mild assumptions. We omit the proof of Lemmas D.1 and D.2 as it can be obtained similarly by setting $b = 0$ and tracking the changes.

Lemma D.1 (Counterpart of Lemma B.2)**.** *Suppose Assumptions A-B hold. Further, assume $\tau_i \geq 1$ and $\alpha_i = \alpha/\tau_i, \forall i \in \mathcal{S}$ for some positive constant α. Then,* FEDOUT *guarantees:*

$$\begin{aligned}\mathbb{E}\left[f(\boldsymbol{x}^+)\right] - \mathbb{E}\left[f(\boldsymbol{x})\right] \leq -\frac{\alpha}{2}(1-\alpha L_f)\mathbb{E}\left[\|\frac{1}{m}\sum_{i=1}^m \frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+)\|^2\right] - \frac{\alpha}{2}\mathbb{E}\left[\|\nabla f(\boldsymbol{x})\|^2\right] \\ + \frac{\alpha L_f^2}{2m}\sum_{i=1}^m \frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\mathbb{E}\left[\|\boldsymbol{x}_{i,\ell}-\boldsymbol{x}\|^2\right]\Big) + \frac{\alpha^2 L_f}{2}\sigma_f^2.\end{aligned} \tag{68}$$

Lemma D.2 (Counterpart of Lemma B.4)**.** *Suppose Assumptions A-B hold. Further, assume $\tau_i \geq 1$ and $\alpha_i = \alpha/\tau_i, \forall i \in \mathcal{S}$, where $\alpha \leq 1/(4M_f)$. Then, $\forall \ell \in \{0,\ldots,\tau_i - 1\}$,* FEDOUT *guarantees:*

$$\mathbb{E}\left[\|\boldsymbol{x}_{i,\ell}-\boldsymbol{x}\|^2\right] \leq c_1\tau_i^2\alpha_i^2\mathbb{E}\Big[\|\nabla f(\boldsymbol{x})\|^2\Big] + c_2\tau_i^2\alpha_i^2\sigma_f^2. \tag{69}$$

for some constant c_1 and c_2.

D.1. Proof of Theorem 3.2

Proof. Combining Lemmas D.1 and D.2, we obtain

$$\begin{aligned}\mathbb{E}\left[f(\boldsymbol{x}^{k+1})\right] - \mathbb{E}\left[f(\boldsymbol{x}^k)\right] &\leq -\frac{\alpha_k}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] + \frac{L_f^2\alpha_k}{2m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\mathbb{E}\Big[\|\boldsymbol{x}_{i,\ell}^k-\boldsymbol{x}^k\|^2\Big] \\ &\quad - \frac{\alpha_k}{2}(1-\alpha_k L_f)\mathbb{E}\left[\|\frac{1}{m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\bar{\boldsymbol{h}}_i(\boldsymbol{x}_{i,\ell}^k,\boldsymbol{y}^{k+1})\|^2\right] + \frac{\alpha_k^2 L_f}{2}\sigma_f^2 \\ &\leq -\frac{\alpha_k}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] + \frac{L_f^2\alpha_k}{2m}\sum_{i=1}^m\frac{1}{\tau_i}\sum_{\ell=0}^{\tau_i-1}\mathbb{E}\Big[\|\boldsymbol{x}_{i,\ell}^k-\boldsymbol{x}^k\|^2\Big] + \frac{\alpha_k^2 L_f}{2}\sigma_f^2 \\ &\leq -\frac{\alpha_k}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] + \frac{c_1 L_f^2\alpha_k^3}{2}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] + \frac{c_2\alpha_k^3 L_f^2}{2}\sigma_f^2 + \frac{\alpha_k^2 L_f}{2}\sigma_f^2. \\ &\leq -\frac{\alpha_k}{4}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] + (c_2+1)\alpha_k^2 L_f\sigma_f^2.\end{aligned} \tag{70}$$

where the second inequality follows by choosing $\bar{\alpha} = \frac{1}{2(c_1+c_2)}$.

Then telescoping gives

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] \leq \mathcal{O}\left(\frac{1}{K} + \frac{\sigma_f}{\sqrt{K}}\right). \tag{71}$$

□

E. Proof for Federated Compositional Optimization

Note that in the stochastic compositional problem, the inner function $f_i(\boldsymbol{x},\boldsymbol{y};\xi) := f_i(\boldsymbol{y};\xi)$ for all $i \in \mathcal{S}$, and the outer function is $g_i(\boldsymbol{x},\boldsymbol{y};\zeta) := \frac{1}{2}\|\boldsymbol{y} - h_i(\boldsymbol{x};\zeta)\|^2$, for all $i \in \mathcal{S}$. Then, we have

$$\begin{aligned}\min_{\boldsymbol{x}\in\mathbb{R}^{d_1}} \quad & f(\boldsymbol{x}) = \tfrac{1}{m}\textstyle\sum_{i=1}^m f_i\left(\boldsymbol{y}^*(\boldsymbol{x})\right) \\ \text{subj. to} \quad & \boldsymbol{y}^*(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{y}\in\mathbb{R}^{d_2}} \ \tfrac{1}{m}\textstyle\sum_{i=1}^m g_i\left(\boldsymbol{x},\boldsymbol{y}\right).\end{aligned} \tag{72a}$$

Here,

$$f_i(\boldsymbol{y}^*(\boldsymbol{x})) := \mathbb{E}_{\xi\sim\mathcal{C}_i}[f_i(\boldsymbol{y}^*(\boldsymbol{x});\xi)],$$
$$g_i(\boldsymbol{x},\boldsymbol{y}) := \frac{1}{2}\mathbb{E}_{\zeta\sim\mathcal{D}_i}[\|\boldsymbol{y}-h_i(\boldsymbol{x};\zeta)\|^2] \tag{72b}$$

are the loss functions of the i^{th} client, $(\xi,\zeta)\sim(\mathcal{C}_i,\mathcal{D}_i)$ are the outer/inner sampling distributions for the i^{th} client.

One can notice that $\nabla_y g_i(\boldsymbol{x},\boldsymbol{y}) = h_i(\boldsymbol{x};\zeta)$, $\nabla_{\boldsymbol{yy}} g(\boldsymbol{x},\boldsymbol{y};\zeta) = \mathbf{I}_{d_1\times d_1}$, and $\nabla_{\boldsymbol{xy}} g(\boldsymbol{x},\boldsymbol{y};\zeta) = -\frac{1}{m}\sum_{i=1}^m \nabla h_i(\boldsymbol{x};\zeta)^\top$. Hence, using (72), we have

$$\begin{aligned}\nabla f_i(\boldsymbol{x}) :&= \nabla_{\boldsymbol{x}} f_i\left(\boldsymbol{y}^*(\boldsymbol{x})\right)\\ &\quad - \nabla^2_{\boldsymbol{xy}} g(\boldsymbol{x},\boldsymbol{y}^*(\boldsymbol{x}))[\nabla^2_{\boldsymbol{y}} g(\boldsymbol{x},\boldsymbol{y}^*(\boldsymbol{x}))]^{-1}\nabla_{\boldsymbol{y}} f_i\left(\boldsymbol{y}^*(\boldsymbol{x})\right)\\ &= (\frac{1}{m}\sum_{i=1}^m \nabla h_i(\boldsymbol{x};\zeta))^\top \nabla_{\boldsymbol{y}} f_i(\boldsymbol{y}^\star(\boldsymbol{x});\xi).\end{aligned} \tag{73}$$

In this case, we can obtain an approximate gradient $\nabla f_i(\boldsymbol{x})$ by replacing $\boldsymbol{y}^\star(\boldsymbol{x})$ with $\boldsymbol{y}$; that is $\bar{\nabla} f_i(\boldsymbol{x},\boldsymbol{y}) = (\frac{1}{m}\sum_{i=1}^m \nabla h_i(\boldsymbol{x};\zeta))^\top \nabla_{\boldsymbol{y}} f_i(\boldsymbol{y};\xi)$. It should be mentioned that in the compositional case $b=0$. Thus, we can apply FEDNEST using the above gradient approximations.

E.1. Supporting Lemmas

Let $\boldsymbol{z}=(\boldsymbol{x},\boldsymbol{y})\in\mathbb{R}^{d_1+d_2}$. We make the following assumptions that are counterparts of Assumptions A and B.

Assumption G. *For all $i\in[m]$:*

(E1) *$f_i(\boldsymbol{z}), \nabla f_i(\boldsymbol{z}), h_i(\boldsymbol{z}), \nabla h_i(\boldsymbol{z})$ are respectively $\ell_{f,0}$, $\ell_{f,1}, \ell_{h,0}, \ell_{h,1}$-Lipschitz continuous; and*

(E2) *$f_i(\boldsymbol{x},\boldsymbol{y})$ is μ_f-strongly convex in $\boldsymbol{y}$ for any fixed $\boldsymbol{x}\in\mathbb{R}^{d_1}$.*

In this case, we have the condition number $\kappa=1$.

Assumption H. *For all $i\in[m]$:*

(F1) *$\nabla f_i(\boldsymbol{z};\xi)$, $h_i(\boldsymbol{x};\zeta)$, $\nabla h_i(\boldsymbol{x};\zeta)$ are unbiased estimators of $\nabla f_i(\boldsymbol{z})$, $\nabla h_i(\boldsymbol{x})$, and $\nabla h_i(\boldsymbol{x})$.*

(F2) *Their variances are bounded, i.e., $\mathbb{E}_\xi[\|\nabla f_i(\boldsymbol{z};\xi)-\nabla f_i(\boldsymbol{z})\|^2]\le\sigma_f^2$, $\mathbb{E}_\zeta[\|h_i(\boldsymbol{z};\zeta)-h_i(\boldsymbol{z})\|^2]\le\sigma_{h,0}^2$, and $\mathbb{E}_\zeta[\|\nabla h_i(\boldsymbol{z};\zeta)-\nabla h_i(\boldsymbol{z})\|^2]\le\sigma_{h,1}^2$ for some $\sigma_f^2,\sigma_{h,0}^2$, and $\sigma_{h,1}^2$.*

Lemma E.1. *Under Assumption G, we have*

$$\|\bar{\nabla} f_i(\boldsymbol{x},\boldsymbol{y})-\nabla f_i(\boldsymbol{x},\boldsymbol{y}^*(\boldsymbol{x}))\|\le\bar{L}_f\|\boldsymbol{y}^*(\boldsymbol{x})-\boldsymbol{y}\|,\ \text{for all } i\in\{1,\dots,m\}, \tag{74a}$$
$$\|\nabla f(\boldsymbol{x}_1)-\nabla f(\boldsymbol{x}_2)\|\le L_f\|\boldsymbol{x}_1-\boldsymbol{x}_2\|, \tag{74b}$$
$$\|\boldsymbol{y}^*(\boldsymbol{x}_1)-\boldsymbol{y}^*(\boldsymbol{x}_2)\|\le L_y\|\boldsymbol{x}_1-\boldsymbol{x}_2\|, \tag{74c}$$
$$\|\nabla\boldsymbol{y}^*(\boldsymbol{x}_1)-\nabla\boldsymbol{y}^*(\boldsymbol{x}_2)\|\le L_{yx}\|\boldsymbol{x}_1-\boldsymbol{x}_2\|, \tag{74d}$$
$$\mathbb{E}\left[\|\bar{\boldsymbol{h}}_i(\boldsymbol{x},\boldsymbol{y})-\boldsymbol{h}_i(\boldsymbol{x},\boldsymbol{y})\|^2\right]\le\tilde{\sigma}_f^2, \tag{74e}$$
$$\mathbb{E}\left[\|\boldsymbol{h}_i(\boldsymbol{x}_{i,\ell},\boldsymbol{y}^+)\|^2|\mathcal{F}_{i,\ell-1}\right]\le\tilde{D}_f^2. \tag{74f}$$

Here,

$$M_f=\ell_{h,0}\ell_{f,1},\quad L_y=\ell_{h,0},\quad L_f=\ell_{h,0}^2\ell_{f,1}+\ell_{f,0}\ell_{h,1},\quad L_{yx}=\ell_{h,1}$$
$$\tilde{\sigma}_f^2=\ell_{h,0}^2\sigma_f^2+(\ell_{f,0}^2+\sigma_f^2)\sigma_{h,1}^2,\quad \tilde{D}_f^2=(\ell_{f,0}^2+\sigma_f^2)(\ell_{h,0}^2+\sigma_{h,1}^2).$$

where $\ell_{f,0}$, $\ell_{f,1}$, $\ell_{h,0}$ $\ell_{h,1}$, and μ_f are given in Assumption G.

Stochastic Compositional Optimization

		Non-Federated		
	FEDNEST	**ALSET**	**SCGD**	**NASA**
batch size	$\mathcal{O}(1)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$
samples	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-2})$

Table 2: Sample complexity of FEDNEST and comparable non-federated methods to find an ϵ-stationary point of f. ALSET (Chen et al., 2021a), SCGD (Wang et al., 2017), NASA (Ghadimi et al., 2020).

Corollary E.1 (Compositional). *Under the same conditions as in Theorem 3.1, if we select* $T = 1, \bar{\alpha} = 1, \eta = \frac{1}{L_{yx}}$ *in* (11). *Then,*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right). \tag{75}$$

Proof. Let

$$\begin{aligned}
\bar{\alpha}_1 &:= \frac{1}{2\ell_{h,0}^2\ell_{f,1} + 4\ell_{h,0}\ell_{f,1} + \ell_{h,0}\ell_{f,1}\ell_{h,1}},\\
\bar{\alpha}_2 &:= \frac{1}{8(\mu_g + \ell_{g,1})(40\ell_{h,0}^2\ell_{f,1} + 2\ell_{h,1}\tilde{C}_f^2\bar{\alpha}_1)},\\
\bar{\beta} &:= \frac{1}{4\rho_g}\left(20\ell_{h,0}^2\ell_{f,1} + \tilde{D}_f^2\bar{\alpha}_1\right).
\end{aligned} \tag{76}$$

We select

$$\alpha_k = \min\{\bar{\alpha}_1, \bar{\alpha}_2, \frac{\bar{\alpha}}{\sqrt{K}}\}, \qquad \beta_k = \frac{\bar{\beta}\alpha_k}{T}. \tag{77}$$

Then, since $T = 1, \bar{\alpha} = 1, \eta = \frac{1}{L_{yx}}$, (58) gives

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] \le \mathcal{O}\left(\frac{1}{\sqrt{K}}\right). \tag{78}$$

This completes the proof. □

Corrollary E.1 implies that for the compositional problem, the convergence rate of FEDNEST to the stationary point of f is $\mathcal{O}(1/\sqrt{K})$. This matches the convergence rate of non-federated stochastic algorithms such as ALSET (Chen et al., 2021a), SCGD (Wang et al., 2017), NASA (Ghadimi et al., 2020) (Table 2).

F. Other Technical Lemmas

We collect additional technical lemmas and facts in this section.

Lemma F.1. *For any set of vectors* $\{\boldsymbol{x}_i\}_{i=1}^m$ *with* $\boldsymbol{x}_i \in \mathbb{R}^d$, *we have*

$$\left\|\sum_{i=1}^m \boldsymbol{x}_i\right\|^2 \le m\sum_{i=1}^m \|\boldsymbol{x}_i\|^2. \tag{79}$$

Lemma F.2. *For any* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, *the following holds for any* $c > 0$*:*

$$\|\boldsymbol{x} + \boldsymbol{y}\|^2 \le (1+c)\|\boldsymbol{x}\|^2 + \left(1 + \frac{1}{c}\right)\|\boldsymbol{y}\|^2. \tag{80}$$

Lemma F.3. *For any set of independent, mean zero random variables* $\{\boldsymbol{x}_i\}_{i=1}^m$ *with* $\boldsymbol{x}_i \in \mathbb{R}^d$, *we have*

$$\mathbb{E}\left[\left\|\sum_{i=1}^m \boldsymbol{x}_i\right\|^2\right] = \sum_{i=1}^m \mathbb{E}\left[\|\boldsymbol{x}_i\|^2\right]. \tag{81}$$

The following lemma characterizes the quality of the approximations (10a) and (10b).

Lemma F.4. *Under Assumptions A-B, the approximate inverse Hessian $\widehat{\boldsymbol{H}}_{\boldsymbol{y}}$ satisfies the following for any $\boldsymbol{x}, \boldsymbol{y}$:*

$$\left\|[\nabla_y^2 g(\boldsymbol{x},\boldsymbol{y})]^{-1} - \mathbb{E}_{\mathcal{W}}[\widehat{\boldsymbol{H}}_{\boldsymbol{y}}]\right\| \leq \frac{1}{\mu_g}\big(\frac{\kappa_g - 1}{\kappa_g}\big)^N,$$
$$\mathbb{E}_{\mathcal{W}}\left[\left\|[\nabla_y^2 g(\boldsymbol{x},\boldsymbol{y})]^{-1} - \widehat{\boldsymbol{H}}_{\boldsymbol{y}}\right\|\right] \leq \frac{2}{\mu_g}. \tag{82}$$

where $\mathcal{W} := \{\mathcal{S}_n, \xi_i, \zeta_i, \xi_{i,0}, \zeta_{i,n} \mid i \in \mathcal{S}_n, \quad 0 \leq n \leq N'\}$. Further, for all $i \in \mathcal{S}$:

$$\left\|\mathbb{E}_{\mathcal{W}}[\boldsymbol{h}_i^I(\boldsymbol{x},\boldsymbol{y})] - \bar{\nabla}^I f_i(\boldsymbol{x},\boldsymbol{y})\right\| \leq b, \tag{83}$$

where $b := \frac{\ell_{g,1}\ell_{f,1}}{\mu_g}\big(\frac{\kappa_g - 1}{\kappa_g}\big)^N$.

Proof. The proof follows the idea of (Ghadimi & Wang, 2018). First, note that by independency of N', $\zeta_{i,n}$, and $\mathcal{S}_n$, and under Assumption B, we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{W}}[\widehat{\boldsymbol{H}}_{\boldsymbol{y}}] &= \mathbb{E}_{\mathcal{W}}\left[\frac{N}{\ell_{g,1}}\prod_{n=1}^{N'}\big(\boldsymbol{I} - \frac{1}{\ell_{g,1}|\mathcal{S}_n|}\sum_{i=1}^{|\mathcal{S}_n|}\nabla_{\boldsymbol{y}}^2 g_i(\boldsymbol{x},\boldsymbol{y};\zeta_{i,n})\big)\right] \\
&= \mathbb{E}_{N'}\left[\mathbb{E}_{\mathcal{S}_{1:N'}}\left[\mathbb{E}_{\zeta}\left[\frac{N}{\ell_{g,1}}\prod_{n=1}^{N'}\big(\boldsymbol{I} - \frac{1}{\ell_{g,1}|\mathcal{S}_n|}\sum_{i=1}^{|\mathcal{S}_n|}\nabla_{\boldsymbol{y}}^2 g_i(\boldsymbol{x},\boldsymbol{y};\zeta_{i,n})\big)\right]\right]\right] \\
&= \tfrac{1}{\ell_{g,1}}\sum_{n=0}^{N-1}\left[\boldsymbol{I} - \tfrac{1}{\ell_{g,1}}\nabla_{\boldsymbol{y}}^2 g(\boldsymbol{x},\boldsymbol{y})\right]^n,
\end{aligned} \tag{84}$$

where the last equality follows from the uniform distribution of N'.

Note that since, $\boldsymbol{I} \succeq \frac{1}{\ell_{g,1}}\nabla_{\boldsymbol{y}}^2 g \succeq \kappa_g$ due to Assumptions A, we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{W}}[\|\widehat{\boldsymbol{H}}_{\boldsymbol{y}}\|] &\leq \frac{N}{\ell_{g,1}}\mathbb{E}_{\mathcal{W}}\left[\prod_{n=1}^{N'}\left\|\boldsymbol{I} - \frac{1}{\ell_{g,1}|\mathcal{S}_n|}\sum_{i=1}^{|\mathcal{S}_n|}\nabla_{\boldsymbol{y}}^2 g_i(\boldsymbol{x},\boldsymbol{y};\zeta_{i,n})\right\|^2\right] \\
&= \frac{N}{\ell_{g,1}}\mathbb{E}_{N'}\left[1-\kappa_g\right]^{N'} = \frac{1}{\ell_{g,1}}\sum_{n=0}^{N-1}\left[1-\kappa_g\right]^n \leq \frac{1}{\mu_g}.
\end{aligned}$$

The reminder of the proof is similar to (Ghadimi & Wang, 2018). □

F.1. Poof of Lemma 2.1

Proof. Given $\boldsymbol{x} \in \mathbb{R}^{d_1}$, the optimality condition of the inner problem in (1) is $\nabla_y g(\boldsymbol{x}, y) = 0$. Now, since $\nabla_{\boldsymbol{x}}\left(\nabla_{\boldsymbol{y}} g(\boldsymbol{x},\boldsymbol{y})\right) = 0$, we obtain

$$0 = \sum_{j=1}^{m}\left(\nabla_{xy}^2 g_j\left(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})\right) + \nabla\boldsymbol{y}^*(\boldsymbol{x})\nabla_{yy}^2 g_j\left(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})\right)\right),$$

which implies

$$\nabla\boldsymbol{y}^*(\boldsymbol{x}) = -\left(\sum_{i=1}^{m}\nabla_{xy}^2 g_i\left(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})\right)\right)\left(\sum_{i=1}^{m}\nabla_{yy}^2 g_j(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}))\right)^{-1}.$$

The results follows from a simple application of the chain rule to f as follows:

$$\nabla f\left(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})\right) = \nabla_{\boldsymbol{x}} f\left(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})\right) + \nabla\boldsymbol{y}^*(\boldsymbol{x})\nabla_y f\left(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})\right).$$

□

	definition		properties			
	outer optimization	inner optimization	global outer gradient	global IHGP	global inner gradient	#comm. rounds
FEDNEST	Alg. 2	Alg. 4	yes	yes	yes	>2
LFEDNEST	Alg. 5	Alg. 6	no	no	no	2

Table 3: Definition of studied algorithms by used local optimization algorithms and server updates and resulting properties of these algorithms.

G. LFEDNEST

Implementing FEDINN and FEDOUT naively by using the global direct and indirect gradients and sending the local information to the server that would then calculate the global gradients leads to a communication and space complexity of which can be prohibitive for large-sized d_1 and d_2.

One can consider possible local variants of FEDINN and FEDOUT tailore to such scenarios. Each of the possible algorithms (See Table 3) can then either use the global gradient or only the local gradient, either use a SVRG or SGD.

Algorithm 5 $x^+ = \textbf{LFEDOUT}\,(x, y, \alpha)$ for stochastic bilevel, min-max, and compositional problems

$x_{i,0} = x$ and $\alpha_i \in (0, \alpha]$
for $i \in \mathcal{S}$ **in parallel do**
  for $\ell = 0, \dots, \tau_i - 1$ **do**
    $h_{i,\ell} = \nabla_x f_i(x_{i,\ell}, y; \xi_{i,\ell}) - \nabla^2_{xy} g_i(x_{i,\ell}, y; \zeta_{i,\ell}) \prod_{n=1}^{N'} \left(I - \frac{1}{\ell_{g,1}} \nabla^2_y g_i(x_{i,\ell}, y; \zeta_{i,n})\right) \nabla_x f_i(y_{i,\ell}, y, \xi_{i,\ell})$
    $h_{i,\ell} = \nabla_x f_i(x_{i,\ell}, y; \xi_{i,\ell})$
    $h_{i,\ell} = \nabla h_i(x_{i,\ell}; \zeta_{i,\ell}) \nabla f_i(y_{i,\ell}; \xi_{i,\ell})$
    $x_{i,\ell+1} = x_{i,\ell} - \alpha_i h_{i,\ell}$
  end for
end for
$x^+ = |\mathcal{S}|^{-1} \sum_{i \in \mathcal{S}} x_{i,\tau_i}$

Algorithm 6 $y^+ = \textbf{LFEDINN}\,(x, y, \beta)$

$\mathbb{G}_i(\cdot) \leftarrow \nabla_y g_i(x, \cdot)$ (bilevel), $-h_i(\cdot)$ (composite), $-\nabla_y f_i(x, \cdot)$ (min-max)
$y_{i,0} = y$ and $\beta_i \in (0, \beta]$
for $i \in \mathcal{S}$ **in parallel do**
  for $\ell = 0, \dots, \tau_i - 1$ **do**
    $q_{i,\ell} = \mathbb{G}_i(y_{i,\ell}; \zeta_{i,\ell})$
    $y_{i,\ell+1} = y_{i,\ell} - \beta_i q_{i,\ell}$
  end for
end for
$y^+ = |\mathcal{S}|^{-1} \sum_{i \in \mathcal{S}} y_{i,\tau_i}$