Secure Bilevel Asynchronous Vertical Federated Learning with Backward Updating

Qingsong Zhang\textsuperscript{1,2}, Bin Gu\textsuperscript{3,4}, Cheng Deng\textsuperscript{1,*}, and Heng Huang\textsuperscript{3,5,*}

\textsuperscript{1}School of Electronic Engineering, Xidian University, Xi’an 710071, China; \textsuperscript{2}JD Tech, Beijing 100176, China.
\textsuperscript{3}JD Finance America Corporation, Mountain View, CA, USA; \textsuperscript{4}MBZUAI, United Arab Emirates
\textsuperscript{5}Electrical and Computer Engineering, University of Pittsburgh, PA, USA
\{qszhang1995, jsgubin, chdeng.xd, henghuanghh\}@gmail.com

Abstract

Vertical federated learning (VFL) attracts increasing attention due to the emerging demands of multi-party collaborative modeling and concerns of privacy leakage. In the real VFL applications, usually only one or partial parties hold labels, which makes it challenging for all parties to collaboratively learn the model without privacy leakage. Meanwhile, most existing VFL algorithms are trapped in the synchronous computations, which leads to inefficiency in their real-world applications. To address these challenging problems, we propose a novel VFL framework integrated with new backward updating mechanism and bilevel asynchronous parallel architecture (VF\textsuperscript{B}\textsuperscript{2}), under which three new algorithms, including VFB\textsuperscript{2}-SGD, -SVRG, and -SAGA, are proposed. We derive the theoretical results of the convergence rates of these three algorithms under both strongly convex and nonconvex conditions. We also prove the security of VFB\textsuperscript{2} under semi-honest threat models. Extensive experiments on benchmark datasets demonstrate that our algorithms are efficient, scalable and lossless.

1 Introduction

Federated learning (McMahan et al. 2016; Smith et al. 2017; Kairouz et al. 2019) has emerged as a paradigm for collaborative modeling with privacy-preserving. A line of recent works (McMahan et al. 2016; Smith et al. 2017) focus on the horizontal federated learning, where each party has a subset of samples with complete features. There are also some works (Gascón et al. 2016; Yang et al. 2019b; Dang et al. 2020) studying the vertical federated learning (VFL), where each party holds a disjoint subset of features for all samples. In this paper, we focus on VFL that has attracted much attention from the academic and industry due to its wide applications to emerging multi-party collaborative modeling with privacy-preserving.

Currently, there are two mainstream methods for VFL, including homomorphic encryption (HE) based methods and exchanging the raw computational results (ERCR) based methods. The HE based methods (Hardy et al. 2017; Cheng et al. 2019) leverage HE techniques to encrypt the raw data and then use the encrypted data (ciphertext) for training model with privacy-preserving. However, there are two major drawbacks of HE based methods. First, the complexity of homomorphic mathematical operation on ciphertext field is very high, thus HE is extremely time consuming for modeling (Liu, Ng, and Zhang 2015; Liu et al. 2019). Second, approximation is required for HE to support operations of non-linear functions, such as Sigmoid and Logarithmic functions, which inevitably causes loss of the accuracy for various machine learning models using non-linear functions (Kim et al. 2018; Yang et al. 2019a). Thus, the inefficiency and inaccuracy of HE based methods dramatically limit their wide applications to realistic VFL tasks.

ERCR based methods (Zhang et al. 2018; Hu et al. 2019; Gu et al. 2020b) leverage labels and the raw intermediate computational results transmitted from the other parties to compute stochastic gradients, and thus use distributed stochastic gradient descent (SGD) methods to train VFL models efficiently. Although ERCR based methods circumvent aforementioned drawbacks of HE based methods, existing ERCR based methods are designed with only considering that all parties have labels, which is not usually the case in real-world VFL tasks. In realistic VFL applications, usually only one or partial parties (denoted as active parties) have the labels, and the other parties (denoted as passive parties) can only provide extra feature data but do not have labels. When these ERCR based methods are applied to the real situation with both active and passive parties, the algorithms even cannot guarantee the convergence because only active parties can update the gradient of loss function based on labels but the passive parties cannot, i.e. partial model parameters are not optimized during the training process. Thus, it comes to the crux of designing the proper algorithm for solving real-world VFL tasks with only one or partial parties holding labels.

Moreover, algorithms using synchronous computation (Gong, Fang, and Guo 2016; Zhang et al. 2018) are inefficient when applied to real-world VFL tasks, especially, when computational resources in the VFL system are unbalanced. Therefore, it is desired to design the efficient asynchronous algorithms for real-world VFL tasks. Although there have been several works studying asynchronous VFL algorithms (Hu et al. 2019; Gu et al. 2020b), it is still an open problem to design asynchronous algorithms for solving real-world VFL tasks with only one or partial parties holding labels.

In this paper, we address these challenging problems by proposing a novel framework (VF\textsuperscript{B}\textsuperscript{2}) integrated with the novel backward updating mechanism (BUM) and bilevel asynchronous parallel architecture (BAPA). Specifically, the
We summarize the contributions of this paper as follows.

- We are the first to propose the novel backward updating mechanism for ERCR based VFL algorithms, which enables all parties, rather than only parties holding labels, to collaboratively learn the model with privacy-preserving and without hampering the accuracy of final model.
- We design a bilevel asynchronous parallel architecture that enables all parties asynchronously update the model through backward updating, which is efficient and scalable.
- We propose three new algorithms for VFL, including VFB\(^2\)-SGD, -SVRG, and -SAGA under VFB\(^2\). Moreover, we theoretically prove their convergence rates for both strongly convex and nonconvex problems.

Notations. \( \bar{w} \) denotes the inconsistent read of \( w \). \( \bar{w} \) denotes \( w \) to compute local stochastic gradient of loss function for collaborators, which maybe stale due to communication delay. \( \psi(t) \) is the corresponding party performing the \( t \)-th global iteration. Given a finite set \( S \), \( |S| \) denotes its cardinality.

### 2 Problem Formulation

Given a training set \( \{x_i, y_i\}_{i=1}^n \), where \( y_i \in \{-1, +1\} \) for binary classification task or \( y_i \in \mathbb{R} \) for regression problem and \( x_i \in \mathbb{R}^d \), we consider the model in a linear form of \( w^\top x \), where \( w \in \mathbb{R}^d \) corresponds to the model parameters. For VFL, \( x_i \) is vertically distributed among \( q \) parties, i.e., \( x_i = [(x_i)_{G_1}; \cdots; (x_i)_{G_q}] \), where \( (x_i)_{G_\ell} \in \mathbb{R}^{d_\ell} \) is stored on the \( \ell \)-th party and \( \sum_{\ell=1}^q d_\ell = d \). Similarly, there is \( w = [w_{G_1}; \cdots; w_{G_q}] \). Particularly, we focus on the following regularized empirical risk minimization problem.

\[
\min_{w \in \mathbb{R}^d} f(w) := \frac{1}{n} \sum_{i=1}^n \mathcal{L}(w^\top x_i, y_i) + \lambda \sum_{\ell=1}^q g(w_{G_\ell}), \tag{P}
\]

where \( w^\top x_i = \sum_{\ell=1}^q w_{G_\ell}^\top (x_i)_{G_\ell} \), \( \mathcal{L} \) denotes the loss function, \( \sum_{\ell=1}^q g(w_{G_\ell}) \) is the regularization term, and \( f : \mathbb{R}^d \to \mathbb{R} \) is smooth and possibly nonconvex. Examples of problem (P) include models for binary classification tasks [Conroy and Sajda, 2012, Wang et al., 2017] and models for regression tasks [Shen et al., 2013, Wang et al., 2019].

In this paper, we introduce two types of parties: active party and passive party, where the former denotes data provider holding labels while the latter does not. Particularly, in our problem setting, there are \( m (1 \leq m \leq q) \) active parties. Each active party can play the role of dominator in model updating by actively launching updates. All parties, including both active and passive parties, passively launching updates play the role of collaborator. To guarantee the model security, only active parties know the form of the loss function. Moreover, we assume that the labels can be shared by all parties. Note that this does not obey our intention that only active parties hold the labels before training. The problem studied in this paper is stated as follows:

Given: Vertically partitioned data \( \{(x_i)_{G_\ell}\}_{\ell=1}^q \) stored in \( q \) parties and the labels only held by active parties.

Learn: A machine learning model \( M \) collaboratively learned by both active and passive parties without leaking privacy.

**Lossless Constraint:** The accuracy of \( M \) must be comparable to that of model \( M' \) learned under non-federated learning.

### 3 VFB\(^2\) Framework

In this section, we propose the novel VFB\(^2\) framework. VFB\(^2\) is composed of three components and its systemic structure is illustrated in Fig. 1a. The details of these components are presented in the following.

The key of designing the proper algorithm for solving real-world VFL tasks with both active and passive parties is to make the passive parties utilize the label information for model training. However, it is challenging to achieve this because direct using the labels held by active parties leads to privacy leakage of the labels without training. To address this challenging problem, we design the BUM with painstaking.

**Backward Updating Mechanism:** The key idea of BUM is to make passive parties indirectly use labels to compute stochastic gradient without directly accessing the raw label data. Specifically, the BUM embeds label \( y_i \) into an intermediate value \( \vartheta := \frac{\partial \mathcal{L}(w^\top x_i, y_i)}{\partial (w^\top x_i)} \). Then \( \vartheta \) and \( i \) are distributed backward to the other parties. Consequently, the passive parties can also compute the stochastic gradient and update the model by using the received \( \vartheta \) and \( i \) (please refer to Algorithms 2 and 3 for details). Fig. 1b depicts the case where \( \vartheta \) is distributed from party 1 to the rest parties. In this case, all parties, rather than only active parties, can collaboratively learn the model without privacy leakage.

For VFL algorithms with BUM, dominated updates in different active parties are performed in distributed-memory parallel, while collaborative updates within a party are performed in shared-memory parallel. The difference of parallelism fashion leads to the challenge of developing a new parallel architecture instead of just directly adopting the existing asynchronous parallel architecture for VFL. To tackle

---

**Algorithm 1 Safe algorithm of obtaining \( w^\top x_i \).**

**Input:** \( \{w_{G_\ell}\}_{\ell=1}^q \) and \( \{(x_i)_{G_\ell}\}_{\ell=1}^q \) allocating at each party, index \( i \).

**Do this in parallel**

1: for \( \ell = 1, \ldots, q \) do
2: Generate a random number \( \delta_{\ell} \) and calculate \( w_{G_\ell}^\top (x_i)_{G_\ell} + \delta_{\ell} \).
3: end for
4: Obtain \( \xi_1 = \sum_{\ell=1}^q w_{G_\ell}^\top (x_i)_{G_\ell} + \delta_{\ell} \) through tree structure \( T_1 \).
5: Obtain \( \xi_2 = \sum_{\ell=1}^q \delta_{\ell} \) through totally different tree structure \( T_2 \neq T_1 \).

Output: \( w^\top x_i = \xi_1 - \xi_2 \)
Bilevel Asynchronous Parallel Architecture: The BAPA includes two levels of parallel architectures, where the upper level denotes the inner-party parallel and the lower one is the intra-party parallel. More specifically, the inner-party parallel denotes distributed-memory parallel between active parties, which enables all active parties to asynchronously launch dominated updates; while the intra-party one denotes the shared-memory parallel of collaborative updates within each party, which enables multiple threads within a specific party to asynchronously perform the collaborative updates. Fig. 1b illustrates the BAPA with m active parties.

Secure Aggregation Strategy: The details are summarized in Algorithm 1. Specifically, at step 2, \( w_{j}(x_{i})_{G_{i}} \) is computed locally on the \( \ell \)-th party to prevent the direct leakage of \( w_{G_{i}} \) and \( (x_{i})_{G_{i}} \). Especially, a random number \( \delta_{i} \) is added to \( w_{j}(x_{i})_{G_{i}} \) to mask the value of \( w_{j}(x_{i})_{G_{i}} \), which can enhance the security during aggregation process. At steps 4 and 5, \( \xi_{1} \) and \( \xi_{2} \) are aggregated through tree structures \( T_{1} \) and \( T_{2} \), respectively. Note that \( T_{2} \) is totally different from \( T_{1} \) that can prevent the random value being removed under threat model 1 (defined in section 6). Finally, value of \( w^{+}x_{i} = \sum_{j=1}^{q}w_{j}(x_{i})_{G_{i}} \) is recovered by removing term \( \sum_{j=1}^{q}\delta_{j} \) from \( \sum_{j=1}^{q}(w_{j}(x_{i})_{G_{i}} + \delta_{j}) \) at the output step. Using such aggregation strategy, \((x_{i})_{G_{i}}\) and \( w_{G_{i}} \) are prevented from leaking during the aggregation.

4 Secure Bilevel Asynchronous VFL Algorithms with Backward Updating

SGD (Bottou 2010) is a popular method for learning machine learning (ML) models. However, it has a poor convergence rate due to the intrinsic variance of stochastic gradient.

Thus, many popular variance reduction techniques have been proposed, including the SVRG, SAGA, SPIDER (Johnson et al. 2013; Defazio, Bach, and Lacoste-Julien 2014), Wang et al. (2019) and their applications to other problems (Huang, Chen, and Huang 2019; Huang et al. 2020); [Dang et al. 2020]; Yang et al. 2020a,b; Li et al. 2020; Wei et al. 2019]. In this section we raise three SGD-type algorithms, i.e., the SGD, SVRG and SAGA, which are the most popular ones among SGD-type methods for the appealing performance in practice. We summarize the detailed steps of VFB\(^2\)-SGD in Algorithms 2 and 3. For VFB\(^2\)-SVRG and -SAGA, one just needs to replace the update rule with corresponding one.

As shown in Algorithm 2, at each dominated update, the dominator (an active party) calculates \( \vartheta \) and then distributes \( \vartheta \) to the collaborators (the rest \( q - 1 \) parties). As shown in algorithm 3 for party \( \ell \), once it has received the \( \vartheta \) and \( i \), it will launch a new collaborative update asynchronously. As the dominator, it computes the local stochastic gradient as \( \nabla_{G_{i}}f_{\vartheta}(\hat{w}) = \nabla_{G_{i}}L(\hat{w}) + \lambda \nabla g(\hat{w}_{G_{i}}) \). While, for the collaborator, it uses the received \( \vartheta \) to compute \( \nabla_{G_{i}}L \) and local \( \hat{w} \) to compute \( \nabla_{G_{i}}g \) as shown at step 3 in Algorithm 3. Note that active parties also need perform Algo-
Algorithm 3 VFB²-SGD for the ℓ-th party to passively launch collaborative updates.

**Input:** Local data \( \{(x_i, y_i)\}_{i=1}^{\ell} \) stored on the ℓ-th party, learning rate \( \gamma \).
1. Initialize the necessary parameters (for passive parties).
2. Receive \( \vartheta \) and the index \( i \) from the dominator.
3. Compute \( \hat{\vartheta} = \nabla g_i \vartheta \) \( + \lambda \nabla g_i (\tilde{w}) = \nabla \cdot (x_i, y_i) \).
4. Update \( w_{\vartheta} \leftarrow w_{\vartheta} - \gamma \hat{\vartheta} \).
5. **End parallel**

rithm to collaborate with other dominators to ensure that the model parameters of all parties are updated.

5 Theoretical Analysis

In this section, we provide the convergence analyses. Please see the arXiv version for more details. We first present preliminaries for strongly convex and nonconvex problems.

**Assumption 1.** For \( f_i(w) \) in problem \( \mathcal{P} \) we assume the following conditions hold:
1. **Lipschitz Gradient:** Each function \( f_i, i = 1, \ldots, n \), there exists \( L > 0 \) such that for all \( w, w' \in \mathbb{R}^d \), there is
   \[
   \| \nabla f_i(w) - \nabla f_i(w') \| \leq L \| w - w' \|.  
   \]
2. **Block-Coordinate Lipschitz Gradient:** For \( i = 1, \ldots, n \), there exists an \( L_{\ell} > 0 \) for the \( \ell \)-th block \( g_i \), where \( \ell = 1, \ldots, q \) such that
   \[
   \| \nabla g_i f_i(w + U_{\ell} \Delta \ell) - \nabla g_i f_i(w) \| \leq L_{\ell} \Delta \ell, \]
   where \( \Delta \ell \in \mathbb{R}^{d_{\ell}}, U_{\ell} \in \mathbb{R}^{d \times d_{\ell}} \) and \( [U_1, \ldots, U_q] = I_d \).
3. **Bounded Block-Coordinate Lipschitz Gradient:** There exists a constant \( G \) such that for \( f_i, i = 1, \ldots, n \) and block \( g_i \), \( \ell = 1, \ldots, q \), it holds that \( \| \nabla g_i f_i(w) \|^2 \leq G \).

**Assumption 2.** The regularization term \( q \) is \( L_g \)-smooth, which means that there exists an \( L_g > 0 \) for \( \ell = 1, \ldots, q \) such that for all \( w, w' \in \mathbb{R}^d \), there is
   \[
   \| \nabla g(w_{\ell}) - \nabla g(w_{\ell}') \| \leq L_g \| w_{\ell} - w_{\ell}' \|.  
   \]

Assumption 2 imposes the smoothness on \( g \), which is necessary for the convergence analyses. Because, as for a specific collaborator, it uses the received \( \tilde{w} \) (denoted as \( w \)) to compute \( \nabla g_i \mathcal{L} \) and local \( \tilde{w} \) to compute \( \nabla g_i g = \nabla g(w_{\ell}) \), which makes it necessary to track the behavior of \( g_i \) individually. Similar to previous research works ([Lian et al. 2015], [Huang 2017], Leblond, Pedregosa, and Lacoste-Julien 2017), we introduce the bounded delay as follows.

**Assumption 3.** Bounded Delay: Time delays of inconsistent reading and communication between dominator and its collaborators are upper bounded by \( \tau_1 \) and \( \tau_2 \), respectively.

Given \( \tilde{w} \) as the consistent read of \( w \), which is used to compute the stochastic gradient in dominated updates, following the analysis in ([Gu et al. 2020b]), we have
   \[
   \tilde{w}_t - w_t = \gamma \sum_{d \in D(t)} U_\psi(u) \tilde{\psi}(u),  
   \]
   where \( D(t) = \{ t - 1, \ldots, t - \tau_0 \} \) is a subset of non-overlapped previous iterations with \( \tau_0 \leq \tau_2 \). Given \( \tilde{w} \) as the parameter used to compute the \( \nabla g_i \mathcal{L} \) in collaborative updates, which is the stale state of \( \tilde{w} \) due to the communication delay between the specific dominator and its corresponding collaborators. Then, following the analyses in ([Huo and Huang 2017]), there is
   \[
   \tilde{w}_t = \tilde{w}_{t-\tau_0} + \tilde{w}_t + \gamma \sum_{i' \in D'(t)} U_{\psi(v')} \tilde{\psi}(v'),  
   \]
   where \( D'(t) = \{ t - 1, \ldots, t - \tau_0 \} \) is a subset of previous iterations performed during the communication and \( \tau_0 \leq \tau_2 \).

**Convergence Analysis for Strongly Convex Problem**

**Assumption 4.** Each function \( f_i, i = 1, \ldots, n \), is \( \mu \)-strongly convex, i.e., \( \forall w, w' \in \mathbb{R}^d \) there exists a \( \mu > 0 \) such that
   \[
   f_i(w) \geq f_i(w') + \langle \nabla f_i(w'), w - w' \rangle + \frac{\mu}{2} \| w - w' \|^2.  
   \]

For strongly convex problem, we introduce notation \( K(t) \) that denotes a minimum set of successive iterations fully visiting all coordinates from global iteration number \( t \). Note that this is necessary for the asynchronous convergence analyses of the global model. Moreover, we assume that the size of \( K(t) \) is upper bounded by \( \eta \), i.e., \( |K(t)| \leq \eta \). Based on \( K(t) \), we introduce the epoch number \( v(t) \) as follow.

**Definition 1.** Let \( P(t) \) be a partition of \( \{0, 1, \cdots, t - \sigma' \} \), where \( \sigma' \geq 0 \). For any \( \kappa \leq P(t) \) we have that there exists \( t' \leq t \) such that \( K(t') = \kappa \), and \( v_1 \subseteq P(t) \) such that \( K(0) = \kappa_1 \). The epoch number for the \( t \)-th global iteration, i.e., \( v(t) \), is defined as the maximum cardinality of \( P(t) \).

Given the definition of epoch number \( v(t) \), we have the following theoretical results for \( \mu \)-strongly convex problem.

**Theorem 1.** Under Assumptions 1, 2, and 3 to achieve the accuracy \( \epsilon \) of problem \( \mathcal{P} \) for VFB²-SGD, i.e., \( \mathbb{E}[f(w_{t+1}) - f(w^*)] \leq \epsilon \), it satisfies \( \tilde{w}_{t+1} \geq 44 \left( \frac{GL_2^3}{\mu^3} \right) \log \left( \frac{2L_2^2}{\epsilon} \right) \), the epoch number \( v(t) \) should satisfy
   \[
   v(t) \geq 44 \left( \frac{GL_2^3}{\mu^3} \right) \log \left( \frac{2L_2^2}{\epsilon} \right),  
   \]
   where \( L_* = \max \{ L_i | i \in [1, q] \} \), \( \tau = \max \{ \tau_1, \tau_2, \eta \} \), and \( w^* \) denote the initial point and optimal point, respectively.

**Theorem 2.** Under Assumptions 1, 2, and 3 to achieve the accuracy \( \epsilon \) of problem \( \mathcal{P} \) for VFB²-SVRG, let \( C = (L_*^2 + L_*)^2 \) and \( \rho = \frac{\lambda_\gamma C}{\mu} \), we can carefully choose \( \gamma \) such that
   \[
   1) 1 - 2L_*^2 \gamma^2 \tau > 0; \ 2) \rho > 0; \ 3) \frac{8L_*^2 + 1/2C}{\rho} \leq 0.05;
   \]
   \[
   4) L_*^2 \gamma^2 + 3/2(2C + 5\gamma) \frac{2\lambda_\gamma C}{\rho} \leq \frac{\epsilon}{83},  
   \]
   where \( \lambda_\gamma = \frac{10}{\log(0.25 \log(1 - \rho))} \), the inner epoch number \( v(t) \) should satisfy \( v(t) \geq \frac{100}{\log(0.25 \log(1 - \rho))} \), and the outer loop number \( S \) should satisfy \( S \geq \frac{1}{\log(0.25 \log(1 - \rho))} \).
Theorem 3. Under Assumptions 2 and 5 to achieve the accuracy $\epsilon$ of problem $P$ for $VFB^2$-SAGA, let $c_0 = (2\gamma^3/2 + (L^*\gamma^2 \tau + L_\gamma \gamma^2)\gamma^{3/2} + 8\gamma^2 \tau)^{18GL^2/1-72L^*\gamma^2\tau}$, $c_1 = 2L^*_\tau (L^2_\gamma \gamma^3 \tau + L_\gamma \gamma^2)$, $c_2 = 4(L^2_\gamma \gamma^3 \tau + L_\gamma \gamma^2) L^*_\tau / n$, and $\rho \in (1 - \frac{1}{n}, 1)$, we can choose $\gamma$ such that

1) $1 - 72L^*_\tau \gamma^2 \tau > 0$; 2) $0 < 1 - \frac{\gamma \mu}{4} < 1$;
3) $\gamma \mu (1 - \rho) \left(\frac{\gamma^2}{4} - 2c_1 - c_2\right) \leq \frac{\epsilon}{2}$;
4) $-\frac{\gamma \mu^2}{4} + 2c_1 + c_2 \left(1 + \left(1 - \frac{1}{\rho}\right)(\tau - 1)^{-1}\right) \leq 0$;
5) $-\frac{\gamma \mu^2}{4} + c_2 + c_1 \left(2 + \left(1 - \frac{1}{\rho}\right)(\tau - 1)^{-1}\right) \leq 0$.

The epoch number $v(t)$ should satisfy $v(t) \geq \frac{1}{\log \pi} \log \frac{2(2\tau - 1)}{(\rho + 1/\rho)}\left(\frac{\alpha^2}{4} - 2c_1 - c_2\right)$.

Remark 1. For strongly convex problems, given the assumptions and parameters in corresponding theorems, the convergence rate of $VFB^2$-SGD is $O(1/\log(1/\pi))$, and those of $VFB^2$-SVRG and $VFB^2$-SAGA are $O(1/\log(1/\pi))$.

Convergence Analysis for Nonconvex Problem

Assumption 5. Nonconvex function $f(w)$ is bounded below,

$$f^* := \inf_{w \in \mathbb{R}^d} f(w) > -\infty. \tag{9}$$

Assumption 5 guarantees the feasibility of nonconvex problem (P). For nonconvex problem, we introduce the notation $K'(t)$ that denotes a set of $q$ iterations fully visiting all coordinates, i.e., $K'(t) = \{0, t, t_1, \cdots, t_{\ell_q - 1}\} = \{0, \cdots, q\}$, where the $t_{\ell_q - 1}$ global iteration denotes a dominated update. Moreover, these iterations are performed respectively on a dominator and $q - 1$ different collaborators receiving $\vartheta$ calculated at the $t$-th global iteration. Further, we assume that $K'(t)$ can be completed in $n\gamma$ global iterations, i.e., for all $t' \in A(t)$, there is $\gamma \geq \max\{u|u \in K'(t')\} - t'$. That is, different from $K(t)$, there is $|K'(t)| = q$ and the definition of $K'(t)$ does not emphasize on "successive iterations" due to the difference of analysis techniques between strongly convex and nonconvex problems. Based on $K'(t)$, we introduce the epoch number $v'(t)$ as follow.

Definition 2. $A(t)$ denotes a set of global iterations, where for all $t' \in A(t)$ there is the $t'$-th global iteration denoting a dominated update and $\bigcup_{t' \in A(t)} K'(t') = \{0, 1, \cdots, t\}$. The epoch number $v'(t)$ is defined as $|A(t)|$.

We give the definition of epoch number $v'(t)$, we have the following theoretical results for nonconvex problem.

Theorem 4. Under Assumptions 2 and 5 to achieve the $\epsilon$-first-order stationary point of problem $P$, i.e., $E[\|\nabla f(w)\|] \leq \epsilon$ for stochastic variable $w$, for $VFB^2$-SGD, let $\gamma = \frac{\mu}{4n\gamma c_1}$, if

$$\tau \leq \frac{512\mu \gamma}{\epsilon^2},$$

then the total epoch number $T$ should satisfy

$$T \geq \frac{E[f(w^0) - f^*]}{\epsilon^2} L_* q G,$$

where $L_* = \max\{L, \{L_t\}_{t=1}^n, L_g\}$, $\tau = \max\{\tau_1^*, \tau_2^*, \eta_2^*\}$, $f(w^0)$ is the initial function value and $f^*$ is defined in Eq. (9).

Theorem 5. Under Assumptions 2 and 5 to solve problem $P$ with $VFB^2$-SVRG, let $\gamma = \frac{\mu}{L_\gamma n\gamma^2}$, where $0 < \mu_0 < \frac{1}{\gamma}$, $0 < \alpha \leq 1$, if epoch number $N$ in an outer loop satisfies

$$N \leq \left\lfloor \frac{n^2}{2m_0} \right\rfloor,$$

and $\tau < \min\{\frac{2n^2}{20\mu_0}, \frac{1 - 8\mu_0}{40m_0}\}$, there is

$$1 \leq \frac{1}{T} \sum_{t=1}^{N-1} \sum_{s=0}^{t} \mathbb{E}[\|\nabla f(w_{t,s})\|^2] \leq \frac{L_* n\gamma E[f(w_0) - f^*]}{T \tau}, \tag{11}$$

where $T$ is the total number of epochs, $t_0$ is the start iteration of epoch $t$, $\tau$ is a small value independent of $n$.

Theorem 6. Under Assumptions 2 and 5 to solve problem $P$ with $VFB^2$-SAGA, let $\gamma = \frac{\mu}{L_\gamma n\gamma^2}$, where $0 < \mu_0 < \frac{1}{\gamma}$, $0 < \alpha \leq 1$, if epoch number $T$ satisfies $T \leq \left\lfloor \frac{n^2}{4m_0} \right\rfloor$, and $\tau < \min\{\frac{n^2}{20\mu_0}, \frac{1 - 2\mu_0}{40m_0}\}$, there is

$$1 \leq \frac{1}{T} \sum_{t=0}^{N-1} \mathbb{E}[\|\nabla f(w_{t})\|^2] \leq \frac{L_* n\gamma E[f(w_0) - f^*]}{T \tau}. \tag{12}$$

Remark 2. For nonconvex problems, given conditions in the theorems, the convergence rate of $VFB^2$-SGD is $O(1/\sqrt{T})$, and those of $VFB^2$-SVRG and $VFB^2$-SAGA are $O(1/T)$.

6 Security Analysis

We discuss the data security and model security of $VFB^2$ under two semi-honest threat models commonly used in security analysis (Cheng et al. 2019; Xu et al. 2019; Gu et al. 2020). Specially, these two threat models have different threat abilities, where threat model 2 allows collusion between parties while threat model 1 does not.

- Honest-but-curious (threat model 1): All workers will follow the algorithm to perform the correct computations. However, they may use their own retained records of the intermediate computation result to infer other worker’s data and model.
- Honest-but-colluding (threat model 2): All workers will follow the algorithm to perform the correct computations. However, some workers may collude to infer other worker’s data and model by sharing their retained records of the intermediate computation result.
Similar to [Gu et al. 2020a], we prove the security of VFB² by analyzing and proving its ability to prevent inference attack defined as follows.

**Definition 3 (Inference attack).** An inference attack on the ℓ-th party is to infer \((x_i)_{\ell}^*_i\) (or \(w_{\ell}^*_i\)) belonging to other parties or \(y_i\) hold by active parties without directly accessing them.

**Lemma 1.** Given an equation \(o_i = w_{\ell}^T(x_i)_{\ell} + o_i = \frac{\partial \mathcal{L}(\tilde{w}^T x_i, y_i)}{\partial (w^T x_i)}\) with only \(o_i\) being known, there are infinite different solutions to this equation.

The proof of lemma 1 is shown in the arXiv version. Based on lemma 1, we obtain the following theorem.

**Theorem 7.** Under two semi-honest threat models, VFB² can prevent the inference attack.

**Feature and model security:** During the aggregation, the value of \(o_i = w_{\ell}^T(x_i)_{\ell} + \delta_i\) is transmitted. Under threat model 1, one even can not access the true value of \(o_i\), let alone using relation \(o_i = w_{\ell}^T(x_i)_{\ell} + \delta_i\) to refer \(w_{\ell}^T(x_i)_{\ell}\). Thus, the aggregation process can prevent inference attack under two semi-honest threat models.

**Label security:** When analyze the label of feature, we do not consider the collusion between active parties and passive parties, which will make preventing labels from leaking meaningless. In the backward updating process, if a passive party \(\ell\) wants to infer \(y_i\) through the received \(\vartheta\), it must solve the equation \(\vartheta = \frac{\partial \mathcal{L}(\tilde{w}^T x_i, y_i)}{\partial (w^T x_i)}\). However, only \(\vartheta\) is known to party \(\ell\). Thus, following from lemma 1, we have that it is impossible to exactly infer the labels. Moreover, the collusion between passive parties has no threats to the security of labels. Therefore, the backward updating can prevent inference attack under two semi-honest threat models.

From above analyses, we have that the feature security, label security and model security are guaranteed in VFB².

### 7 Experiments

In this section, extensive experiments are conducted to demonstrate the efficiency, scalability and losslessness of our algorithms. More experiments are presented in the arXiv version.

**Experiment Settings:** All experiments are implemented on a machine with four sockets, and each socket has 12 cores. To simulate the environment with multiple machines (or parties), we arrange an extra thread for each party to schedule its \(k\) threads and support communication with (threads of) the other parties. We use MPI to implement the communication scheme. The data are partitioned horizontally and randomly into \(q\) non-overlapped parts with nearly equal number of features. The number of threads within each party, i.e. \(k\), is set as \(m\). We use the training dataset or randomly select 80% samples as the training data, and the testing dataset or the rest as the testing data. An optimal learning rate \(\gamma\) is chosen from \(\{5e^{-1}, 1e^{-1}, 5e^{-2}, 1e^{-2}, \ldots\}\) with regularization coefficient \(\lambda = 1e^{-4}\) for all experiments.

**Datasets:** We use four classification datasets summarized in Table 1 for evaluation. Especially, \(D_1\) (UCICreditCard) and \(D_2\) (GiveMeSomeCredit) are the real financial datasets from the Kaggle website [https://www.kaggle.com/datasets](https://www.kaggle.com/datasets) which can be used to demonstrate the ability to address real-world tasks; \(D_3\) (new20) and \(D_4\) (webspam) are the large-scale ones from the LIBSVM [Chang and Lin 2011](https://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets/) website. Note that we apply one-hot encoding to categorical features of \(D_1\) and \(D_2\), thus the number of features become 90 and 92, respectively.

**Problems:** We consider \(\ell_2\)-norm regularized logistic regression problem for \(\mu\)-strong convex case

\[
\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-y_i w^T x_i}) + \lambda \frac{1}{2} \|w\|^2, \quad (13)
\]

and the nonconvex logistic regression problem

\[
\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-y_i w^T x_i}) + \lambda \frac{1}{2} \sum_{i=1}^{d} w_i^2. \quad (13)
\]

#### Evaluations of Asynchronous Efficiency and Scalability

To demonstrate the asynchronous efficiency, we introduce the synchronous counterparts of our algorithms (i.e., synchronous VFL algorithms with BUM, denoted as VFB) for comparison. When implementing the synchronous algorithms, there is a synthetic straggler party which may be 30% to 50% slower than the faster party to simulate the real application scenario with unbalanced computational resource.

**Asynchronous Efficiency:** In these experiments, we set \(q = 8\), \(m = 3\) and fix the \(\gamma\) for algorithms with a same SGD-type but in different parallel fashions. As shown in Figs. 3 and 4, the loss v.s. run time curves demonstrate that our algorithms consistently outperform their synchronous counterparts regarding the efficiency.

Moreover, from the perspective of loss v.s. epoch number, we have that algorithms based on SVRG and SAGA have the following speedup scalability in terms of the number of total parties \(q\). Given a fixed \(m\), \(q\)-parties speedup is defined as

\[
q\text{-parties speedup} = \frac{\text{Run time of using 1 party}}{\text{Run time of using } q\text{ parties}}, \quad (14)
\]

| Datasets           | Financial | Large-Scale |
|--------------------|-----------|-------------|
|                     | \(D_1\)   | \(D_2\)     | \(D_3\)   | \(D_4\)   |
| #Samples           | 24,000    | 96,257      | 17,996    | 175,000   |
| #Features          | 90        | 92          | 1,355,191 | 16,609,143|

Table 1: Dataset Descriptions.

[https://www.kaggle.com/datasets](https://www.kaggle.com/datasets)
[https://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets/](https://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets/)
To demonstrate the losslessness of our algorithms, we compare VFB$^2$-SVRG with its non-federated (NonF) counterpart (all data are integrated together for modeling) and ERCR based algorithms but without BUM, i.e., AFSVRG-VP proposed in [Gu et al. 2020a]. Especially, AFSVRG-VP also uses distributed SGD method but can not optimize the parameters corresponding to passive parties due to lacking labels. When implementing AFSVRG-VP, we assume that only half parties have labels, i.e., parameters corresponding to the features held by the other parties are not optimized. Each comparison is repeated 10 times with $m = 3$, $q = 8$, and a same stop criterion, e.g., $10^{-5}$ for $D_1$. As shown in Table 2, the accuracy of our algorithms are the same with those of NonF algorithms and are much better than those of AFSVRG-VP, which are consistent to our claims.

**Evaluation of Losslessness**

To demonstrate the losslessness of our algorithms, we compare VFB$^2$-SVRG with its non-federated (NonF) counterpart (all data are integrated together for modeling) and ERCR based algorithms but without BUM, i.e., AFSVRG-VP proposed in [Gu et al. 2020a]. Especially, AFSVRG-VP also uses distributed SGD method but can not optimize the parameters corresponding to passive parties due to lacking labels. When implementing AFSVRG-VP, we assume that only half parties have labels, i.e., parameters corresponding to the features held by the other parties are not optimized. Each comparison is repeated 10 times with $m = 3$, $q = 8$, and a same stop criterion, e.g., $10^{-5}$ for $D_1$. As shown in Table 2, the accuracy of our algorithms are the same with those of NonF algorithms and are much better than those of AFSVRG-VP, which are consistent to our claims.

**8 Conclusion**

In this paper, we proposed a novel backward updating mechanism for the real VFL system where only one or partial parties have labels for training models. Our new algorithms enable all parties, rather than only active parties, to collaboratively update the model and also guarantee the algorithm convergence, which was not held in other recently proposed ERCR based VFL methods under the real-world setting. Moreover, we proposed a bilevel asynchronous parallel architecture to make ERCR based algorithms with backward updating more efficient in real-world tasks. Three practical SGD-type of algorithms were also proposed with theoretical guarantee.
References

Bottou, L. 2010. Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT’2010, 177–186. Springer.

Chang, C.-C.; and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST) 2(3): 27.

Cheng, K.; Fan, T.; Jin, Y.; Liu, Y.; Chen, T.; and Yang, Q. 2019. SecureBoost: A Lossless Federated Learning Framework. arXiv preprint arXiv:1901.08755.

Conroy, B.; and Sajda, P. 2012. Fast, exact model selection and permutation testing for L2-regularized logistic regression. In Artificial Intelligence and Statistics, 246–254.

Dang, Z.; Li, X.; Gu, B.; Deng, C.; and Huang, H. 2020. Large-Scale Nonlinear AUC Maximization via Triply Stochastic Gradients. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Defazio, A.; Bach, F.; and Lacoste-Julien, S. 2014. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In Advances in NIPS, 1646–1654.

Gascón, A.; Schoppmann, P.; Balle, B.; Raykova, M.; Doerner, J.; Zahur, S.; and Evans, D. 2016. Secure Linear Regression on Vertically Partitioned Datasets. IACR Cryptology ePrint Archive 2016: 892.

Gong, Y.; Fang, Y.; and Guo, Y. 2016. Private data analytics on biomedical sensing data via distributed computation. IEEE/ACM transactions on computational biology and bioinformatics 13(3): 431–444.

Gu, B.; Dang, Z.; Li, X.; and Huang, H. 2020a. Federated Doubly Stochastic Kernel Learning for Vertically Partitioned Data. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2483–2493.

Gu, B.; Xu, A.; Deng, C.; and Huang, H. 2020b. Privacy-Preserving Asynchronous Federated Learning Algorithms for Multi-Party Vertically Collaborative Learning. arXiv preprint arXiv:2008.06233.

Hardy, S.; Henecka, W.; Ivey-Law, H.; Nock, R.; Patrini, G.; Smith, G.; and Thorne, B. 2017. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. arXiv preprint arXiv:1711.10677.

Hu, Y.; Niu, D.; Yang, J.; and Zhou, S. 2019. FDML: A collaborative machine learning framework for distributed features. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2232–2240.

Huang, F.; Chen, S.; and Huang, H. 2019. Faster Stochastic Alternating Direction Method of Multipliers for Nonconvex Optimization. In ICML, 2839–2848.

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

Huo, Z.; and Huang, H. 2017. Asynchronous mini-batch gradient descent with variance reduction for non-convex optimization. In Thirty-First AAAI Conference on Artificial Intelligence.

Johnson, R.; and Zhang, T. 2013. Accelerating stochastic gradient descent using predictive variance reduction. In Advances in NIPS, 315–323.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2019. Advances and Open Problems in Federated Learning. arXiv preprint arXiv:1912.04977.

Kim, M.; Song, Y.; Wang, S.; Xia, Y.; and Jiang, X. 2018. Secure logistic regression based on homomorphic encryption: Design and evaluation. JMIR medical informatics 6(2): e19.

Leblond, R.; Pedregosa, F.; and Lacoste-Julien, S. 2017. Asaga: Asynchronous Parallel Saga. In 20th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017.

Li, M.; Deng, C.; Li, T.; Yan, J.; Gao, X.; and Huang, H. 2020. Towards Transferable Targeted Attack. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 641–649.

Lian, X.; Huang, Y.; Li, Y.; and Liu, J. 2015. Asynchronous parallel stochastic gradient for nonconvex optimization. In Advances in Neural Information Processing Systems, 2737–2745.

Liu, F.; Ng, W. K.; and Zhang, W. 2015. Encrypted gradient descent protocol for outsourced data mining. In 2015 IEEE 29th International Conference on Advanced Information Networking and Applications, 339–346. IEEE.

Liu, Y.; Liu, Y.; Liu, Z.; Zhang, J.; Meng, C.; and Zheng, Y. 2019. Federated Forest. arXiv preprint arXiv:1905.10053.

McMahan, H. B.; Moore, E.; Ramage, D.; Hampson, S.; et al. 2016. Communication-efficient learning of deep networks from decentralized data. arXiv preprint arXiv:1602.05629.

Shen, X.; Alam, M.; Fikse, F.; and Rönnegård, L. 2013. A novel generalized ridge regression method for quantitative genetics. Genetics 193(4): 1255–1268.

Smith, V.; Chiang, C.-K.; Sanjabi, M.; and Talwalkar, A. S. 2017. Federated multi-task learning. In Advances in Neural Information Processing Systems, 4424–4434.

Wang, X.; Ma, S.; Goldfarb, D.; and Liu, W. 2017. Stochastic quasi-Newton methods for nonconvex stochastic optimization. SIAM Journal on Optimization 27(2): 927–956.

Wang, Z.; Ji, K.; Zhou, Y.; Liang, Y.; and Tarokh, V. 2019. SpiderBoost and Momentum: Faster Variance Reduction Algorithms. In Advances in NIPS, 2403–2413.

Wei, K.; Yang, M.; Wang, H.; Deng, C.; and Liu, X. 2019. Adversarial Fine-Grained Composition Learning for Unseen Attribute-Object Recognition. In Proceedings of the IEEE International Conference on Computer Vision, 3741–3749.

Xu, R.; Baracaldo, N.; Zhou, Y.; Anwar, A.; and Ludwig, H. 2019. Hybridalpha: An efficient approach for privacy-preserving federated learning. In Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security, 13–23.
Yang, K.; Fan, T.; Chen, T.; Shi, Y.; and Yang, Q. 2019a. A Quasi-Newton Method Based Vertical Federated Learning Framework for Logistic Regression. *arXiv preprint arXiv:1912.00513*.

Yang, M.; Deng, C.; Yan, J.; Liu, X.; and Tao, D. 2020a. Learning Unseen Concepts via Hierarchical Decomposition and Composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10248–10256.

Yang, Q.; Liu, Y.; Chen, T.; and Tong, Y. 2019b. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10(2): 12.

Yang, X.; Deng, C.; Wei, K.; Yan, J.; and Liu, W. 2020b. Adversarial Learning for Robust Deep Clustering. *Advances in Neural Information Processing Systems* 33.

Zhang, G.-D.; Zhao, S.-Y.; Gao, H.; and Li, W.-J. 2018. Feature-Distributed SVRG for High-Dimensional Linear Classification. *arXiv preprint arXiv:1802.03604*.
Supplementary Materials

We present the related supplements in following sections.

A Explanation of the Bilevel Asynchronous Parallel Architecture

When \( m = 1 \), we just need to set the number of threads within each party as 1, then Bilevel Asynchronous Parallel Architecture (BAPA) reduces to a parallel architecture with multiple parties. While, the updates on passive parties rely on the \( \vartheta \) received from the only active party. In this case, the BAPA behaves likely (just behaves likely not the same as) the server-worker distributed-memory architecture in (Huo and Huang [2017]) for there is a communication delay between the active and passive parties. The difference is that in our BAPA with \( m = 1 \) the worker (i.e., passive parties \( \ell \) in our BAPA) passively send the local \( w_{G_{\ell}^T}^+(x_i)_{\ell} \) to the other parties when \( w^+ x_i \) is required instead of just actively sending the local \( w_{G_{\ell}^T}^+(x_i)_{\ell} \) to the only server (i.e., active in our BAPA). When \( m = q \), then all parties hold labels and the BAPA reduces to the general shared-memory parallel architecture from the perspective of analysis.

B Supplements Related to Tree-Structured Communication

The definition and illustration of totally different tree structures

First, we present the definition of significantly different tree structures mentioned at step 5 in Algorithm[1]

Definition 4 (Two significantly different tree structures[Gu et al. (2020a)]. For two tree structures \( T_1 \) and \( T_2 \) on all parties \( \{1, \ldots, q\} \), they are significantly different if there does not exist a subtree \( T_1 \) of \( T_1 \) and a subtree \( T_2 \) of \( T_2 \) whose size are larger than \( l \) and smaller than \( T_1 \) and \( T_2 \), respectively, such that leaf (\( T_1 \)) = leaf (\( T_2 \)).

Then we present an illusion of the totally different tree structures in Fig. 5

![Tree structure](image)

Figure 5: Illustration of tree-structured communication based on two totally different tree structures \( T_1 \) and \( T_2 \).

As depicted in Fig. 5(a), party 1 aggregates values from parties 1 and 2; party 3 aggregates values from parties 3 and 4; and then party 1, i.e., the aggregator, aggregates these two aggregated values from parties 1 and 3. While, as depicted in Fig. 5(b), party 1 aggregates values from parties 1 and 3; party 2 aggregates values from parties 2 and 4; and then the aggregated values are aggregated from parties 1 and 2 to party 1, i.e., the aggregator. From the aggregation process described above, it is easily to conclude that aggregation through such significantly different tree structures can prevent the leakage of the random value \( \delta_\ell \) when there are no collusion between parties.

An example showing collusion between parties

Then we present an example to show that collusion between parties can remove the random value \( \delta_\ell \) added to \( w_{G_{\ell}^T}^+(x_i)_{\ell} \). Assume that \( \{w_{G_{\ell}^T}^+(x_i)_{\ell} \}^{q}_{\ell=1} \) are aggregated through tree structure \( T_1 \) and \( \{\delta_{\ell} \}^{q}_{\ell=1} \) are aggregated through tree structure \( T_2 \).

In this case, party 3 knows the value of \( w_{G_{\ell=4}^T}^+(x_i)_{\ell=4} + \delta_{\ell=4} \) and party 2 knows the value of \( \delta_{\ell=4} \). Then if there is collusion between parties 2 and 3, \( \delta_{\ell=4} \) added to party 4 can be removed from \( w_{G_{\ell=4}^T}^+(x_i)_{\ell=4} + \delta_{\ell=4} \).

Proof of Lemma[1]

Proof. First, we consider the equation \( \alpha_i = w_{G_{\ell}^T}^+(x_i)_{\ell} \) with two cases, including \( d_\ell \geq 2 \) and \( d_\ell = 1 \). For \( \forall d_\ell \geq 2 \), given an arbitrary non-identity orthogonal matrix \( U \in \mathbb{R}^{d_\ell \times d_\ell} \), we have

\[
(w_{G_{\ell}^T}^TU^T)(U(x_i)_{\ell}) = w_{G_{\ell}^T}^TU^T(x_i)_{\ell} = w_{G_{\ell}^T}^+(x_i)_{\ell} = \alpha_i
\]

(15)
From Eq. 15 we have that given an equation \( a_i = w_{G_i}^T(x_i)_{G_i} \) with only \( a_i \) being known, the solutions corresponding to \( w_{G_i} \) and \( (x_i)_{G_i} \) can be represented as \( (w_{G_i}^T U^T) \) and \( (U(x_i)_{G_i}) \), respectively. However, \( U \) can be an arbitrary different non-identity orthogonal matrix, the solutions are thus infinite. If \( d_\ell = 1 \), give an arbitrary real number \( u \neq 1 \), we have

\[
(w_{G_i}^T u) \left( \frac{1}{u} (x_i)_{G_i} \right) = w_{G_i}^T \left( \frac{1}{u} (x_i)_{G_i} \right) = w_{G_i}^T (x_i)_{G_i} = a_i
\]

(16)

Similar to above analysis, we have that the solutions of equation \( a_i = w_{G_i}^T (x_i)_{G_i} \) are infinite when \( d_\ell = 1 \). As for \( a_i = \frac{\partial \mathcal{L}(w^*_\ell x_i, y_i)}{\partial (w^*_\ell x_i)} \), both \( w^*_\ell x_i \) and loss function are unknown, it is thus impossible to exactly infer the \( y_i \). This completes the proof.

\[\square\]

### C Detailed Algorithmic Steps of VFB\(^2\)-SVRG and -SAGA

In the following, we present the detailed algorithmic steps of VFB\(^2\)-SVRG and -SAGA.

**VFB\(^2\)-SVRG**

The proposed VFB\(^2\)-SVRG with an improved convergence rate than VFB\(^2\)-SGD is shown in Algorithms 4 and 5. Different from VFB\(^2\)-SGD directly using the stochastic gradient for updating, VFB\(^2\)-SVRG adopts the variance reduction technique to control the intrinsic variance of stochastic gradient. Algorithm 4 thus computes \( \bar{\nu}^t := \nabla \hat{g}_i f_i (\bar{w}) - \nabla \hat{g}_i f_i (w^*) + \nabla \hat{g}_i f (w^*) \). While for Algorithm 5 there is \( \bar{\nu}^t = \bar{\nu}^1 \cdot (x_i)_{G_i} + \nabla y(\hat{g}_i) - (\bar{\nu}^1 \cdot (x_i)_{G_i} + \nabla y(g(\hat{g}_i))) + \nabla \hat{g}_i f (w^*) \). where \( \nabla \hat{g}_i f_i (w^*) \) is computed as \( (\bar{\nu}^1 \cdot (x_i)_{G_i} + \nabla \hat{g}_i g(\hat{g}_i^0)) \).

#### Algorithm 4 VFB\(^2\)-SVRG for active party \( \ell \) to actively launch dominated update

**Input:** Local data \( \{(x_i)_{G_i}, y_i \}_{i=1}^n \) stored on the \( \ell \)-th party, learning rate \( \gamma \).

1. Initialize \( w_{G_i}^0 \in \mathbb{R}^{d_\ell} \).
2. for \( s = 0, 1, \ldots, S - 1 \) do
3. Compute \( (w^*)^T x_i \) for \( i = 1, \ldots, n \) based on Algorithm 1
4. Compute \( \theta_0 = \frac{\partial \mathcal{L}(w^*)^T x_i}{\partial (w^*)^T x_i} \) for \( i = 1, \ldots, n \) and the full local gradient \( \nabla f_i (w^*) = \frac{1}{n} \sum_{i=1}^n \nabla f_i (w^*) \), and then distribute all \( \theta_0 \) to the rest parties.
5. \( w_{G_i}^t = w_{G_i}^0 \).
6. Keep doing in parallel (distributed-memory parallel for multiple active parties)
7. Pick an index \( i \) randomly from \( \{1, \ldots, n\} \).
8. Compute \( \bar{w}^\top x_i \) based on tree-structured communication.
9. Compute \( \bar{\theta}_1 = \frac{\partial \mathcal{L}(\bar{w}^\top x_i, y_i)}{\partial \bar{w}^\top x_i} \).
10. Snd \( \bar{\theta}_1 \) and index \( i \) to the rest parties.
11. Compute \( \bar{\nu}^t = \nabla \hat{g}_i f_i (\bar{w}) - \nabla \hat{g}_i f_i (w^*) + \nabla \hat{g}_i f (w^*) \).
12. Update \( w_{G_i}^t := w_{G_i}^t - \gamma \bar{\nu}^t \).
13. End parallel
14. \( w_{G_i}^{t+1} = w_{G_i}^t \).
15. end for

**VFB\(^2\)-SAGA**

VFB\(^2\)-SAGA enjoying the same convergence rate with VFB\(^2\)-SVRG is shown in Algorithms 6 and 7. Different from VFB\(^2\)-SVRG using \( w^* \) as the reference gradient, VFB\(^2\)-SAGA uses the average of history gradients stored in a table. In Algorithm 6 there is \( \bar{\nu}^t = \nabla \hat{g}_i f_i (\hat{w}) - \hat{\alpha}_i^t + \frac{1}{n} \sum_{j=1}^n \hat{\alpha}_j^t \). While, in Algorithm 7 there is \( \bar{\nu}^t = \bar{\nu}^1 \cdot (x_i)_{G_i} + \nabla y(\hat{g}_i) - (\bar{\nu}^1 \cdot (x_i)_{G_i} + \nabla y(g(\hat{g}_i))) + \nabla \hat{g}_i f (w^*) \).

### D Additional Experiments on Regression Task

These experiments are conducted on two datasets for regression task: \( D_6 \) (E2006-tsdf) and \( D_9 \) (YearPredictionMSD) from the LIBSVM (Chang and Lin 2011). \( D_6 \) has 16,087 training samples and 150,306 features. \( D_9 \) has 463,715 training samples and 90 features. Moreover, we apply the min-max normalization technique to the target variables \( y \) of \( D_6 \).

**Problems:** We consider \( \ell_2 \)-norm regularized regression problem for \( \mu \)-strong convex case

\[
\min_{w \in \mathbb{R}^d} f(w) := \frac{1}{n} \sum_{i=1}^n (w^\top x_i - y_i)^2 + \frac{\lambda}{2} \|w\|^2.
\]

(17)
present a more clear illusion of the asynchronous scalability in terms of $q$

Algorithm 5 VFB$^2$-SVRG for the $\ell$-th party to passively launch collaborative updates.

**Input:** Local data $D^\ell$ stored on the $\ell$-th party, learning rate $\gamma$.
1: Initialize $w_{G_\ell} \in \mathbb{R}^{d_\ell}$ (only performed on passive parties).
2: for $s = 0, 1, \ldots, S - 1$ do
3:    Receive all $\theta_{i,s}$ from the dominator and use them to compute the full local gradient $\nabla_{G_\ell} f(w^*) = \frac{1}{n} \sum_{i=1}^{n} \nabla_{G_\ell} f_i(w^*)$.
4:    for $i$ do
5:        Compute $\tilde{\vartheta} \leftarrow \vartheta_{1,i} \cdot (x_i)_{G_\ell} + \nabla g(w_{G_\ell}^n)).$
6:        Compute $w_{G_\ell}^{s+1} \leftarrow \vartheta_{1,i} \cdot (x_i)_{G_\ell} + \nabla g(w_{G_\ell}^n)).$
7:    End for
8: end for

Algorithm 6 VFB$^2$-SAGA for active party $\ell$ to actively launch dominated update

**Input:** Local data $\{(x_i)_{G_\ell}, y_i\}_{i=1}^{n}$ stored on the $\ell$-th party, learning rate $\gamma$.
1: Initialize $w_{G_\ell} \in \mathbb{R}^{d_\ell}$.
2: Compute the local gradient $\tilde{\vartheta}^\ell = \nabla_{G_\ell} f_i(\tilde{w})$, for $\forall i \in \{1, \ldots, n\}$ and $\ell = 1, \ldots, q$ through tree-structured communication.
   (this is performed only at the 1-th global iteration)
3: Pick an index $i$ randomly from $1, \ldots, n$.
4: Compute $\tilde{w}_{\ell} = \sum_{i=1}^{n} (\tilde{w})_{G_\ell} (x_i)_{G_\ell}$ based on tree-structured communication.
5: Compute $\vartheta = \frac{\partial L(w, y_i)}{\partial w(x_i)}$.
6: Send $\vartheta$ and index $i$ to collaborators.
7: Compute $\tilde{\vartheta}^\ell = \nabla_{G_\ell} f_i(\tilde{w}) - \tilde{\vartheta}^\ell + \frac{1}{n} \sum_{j=1}^{n} \tilde{\vartheta}^\ell_j$.
8: Update $w_{G_\ell} \leftarrow w_{G_\ell} - \gamma \tilde{\vartheta}^\ell$.
9: Update $\tilde{\vartheta}^\ell \leftarrow \nabla_{G_\ell} f_i(\tilde{w})$.
End parallel

Output: $w_{G_\ell}$

and the robust linear regression for nonconvex problem

$$\min_{w \in \mathbb{R}^d} f(w) := \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i - \langle x_i, w \rangle),$$

(18)

where $\mathcal{L}(x) := \log(\frac{e^x}{2} + 1)$.

**Asynchronous efficiency:** In these experiments, we set $q = 12, m = 2$ and fix the $\gamma$ for algorithms with a same SGD-type but in different parallel fashions. As shown in Fig. 6 the loss v.s. running time curves demonstrate that our algorithms consistently outperform the corresponding synchronous counterparts in terms of the efficiency.

**Evaluations of the losslessness** To demonstrate that our algorithms are lossless, we compare them with the corresponding non-federated (NonF) algorithms, i.e., all data were integrated together for modeling. For datasets without testing data, we split the data set into 5 parts, and use one of them for testing. Moreover, we use the metric root mean square error (RMSE) for evaluation

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y} - y)^2},$$

(19)

where $\hat{y}$ denotes the prediction value and $y$ is the true value. As shown in Table 3 the results of our algorithms are the same with those of NonF algorithms and are much better than those of AFSVRG-VP, which are consistent to our claims.

**Asynchronous scalability in terms of $q$**

W present a more clear illusion of the asynchronous scalability in terms of $q$ shown in Fig. 7.
As shown in the algorithms, we do not globally label the iterates from different parties. While, Globally labeling the iterates: In this section, we present some preliminaries which are helpful for readers to understand the analysis.

\[ \hat{w}^t \leftarrow w^t_{\hat{G}_t} - \gamma \nabla_{\hat{G}_t} g((\hat{w})_{\hat{G}_t}). \]

E Preliminaries for Convergence Analysis (corresponding to line 254 in the manuscript)

In this section, we present some preliminaries which are helpful for readers to understand the analysis.

**Globally labeling the iterates:** As shown in the algorithms, we do not globally label the iterates from different parties. While, how to define the global iteration counter in the convergence analysis. In this paper, we adopt the “after read” labeling strategy (Leblond, Pedregosa, and Lacoste-Julien 2017), where the global iterate counter is updated as one dominator finishes computing \( w^t_i \), or as one collaborator finishes reading local parameters \( (\hat{w}_i)_{\hat{G}_{ix}(i)} \) (this reading operation is performed after having received information from a specific dominator, e.g., step 3 in Algorithm 5). It means that \( \hat{w}_i \) on a specific
dominated parties is the $t + 1$-th fully completed computation of $\tilde{w}^T x_t$ and $(\tilde{w}_t)_{\psi(t)}$ on a collaborative party is the $t + 1$-th fully completed read of $(w_t)_{\psi(t)}$. Importantly, such a labeling strategy guarantees that $i_t$ and $\tilde{w}_t$ are independent (Leblond, Pedregosa, and Lacoste-Julien 2017), which simplifies the convergence analyses, especially for VFB$^2$-SAGA.

**Global updating rule:** Here we introduce the global updating rule as

$$w_{t+1} = w_t - \gamma U_{\psi(t)} \bar{v}_t^{\psi(t)}$$  \hspace{1cm} (20)

where $\bar{v}_t^{\psi(t)}$ has a different definition on different type of roles (dominator or collaborator). Although the definitions of $\bar{v}_t^{\psi(t)}$ are different on different type of roles, we will build uniform analyses for them.

**Relationship between $v_t$ and $\tilde{w}_t$:** For dominators, $\tilde{w}^T x_t = \sum_{t' = 1}^q (\tilde{w})_{\psi(t')} x_{t'}$ is obtained based on Algorithm 1 in an asynchronous parallel fashion, where $\tilde{w}$ denotes $u$ inconsistently read from different data parties. It means that, vector $(\tilde{w}_t)_{\psi(t)}$ (where $t' \neq t$) may be inconsistent to $(w_t)_{\psi(t)}$, i.e., some blocks of $\tilde{w}_t$ are the same with the ones in $w_t$ (e.g., $(w_t)_{\psi(t')} = (\tilde{w}_t)_{\psi(t')}$), but others are different. Thus we introduce a set $D(t)$ in Eq. 4 and the upper bound of its size is introduced in Assumption 3.

**Relationship between $\tilde{w}_t$ and $\tilde{w}_u$:** For a collaborative party, it use $\bar{v}$ received from dominator to compute $\nabla_{\psi(t)} \mathcal{L}_t$ and we donate $\bar{v} \cdot (x_t)_{\psi(t)}$ at global iteration $t$ as $\nabla_{\psi(t)} \mathcal{L}_t(\tilde{w}_t)$. Since there is a communication delay between dominator and collaborators, $\tilde{w}_t$ maybe an old $\tilde{w}_u$ ($u \leq t$). To describe the relation between $\tilde{w}_t$ and $\tilde{w}_u$, we thus introduce a set $D'(t)$ in Eq. 5 (when $u = t$, $D'(t)$ denotes an empty set). Meanwhile, we introduce an upper bound to the communication delay in Assumption 5.

**Introduction of $\bar{v}_t$ and $\bar{v}_t$:** In Algorithms 2 and 3, we have that for a dominator, there is $\bar{v}_t = \bar{v}_t$. While for collaborators, there is $\bar{v}_t = \bar{v}_t = \bar{v}_t + \nabla_{\psi(t)} g(\bar{w})$ which can be rewritten as $\bar{v}_t = \bar{v}_t + \nabla_{\psi(t)} g(\bar{w})$.

### F Convergence Analyses for Strongly Convex problems

**Convergence Analysis of Theorem 1**

**Lemma 2.** For VFB$^2$-SGD, for $\forall t$, there is

$$\mathbb{E}[||\bar{v}_t^{\psi(t)}||^2] \leq \frac{2G}{1 - \lambda_1}$$  \hspace{1cm} (21)

where there is $\lambda_1 = 2L_2^2 \gamma^2 \tau$.

**Proof of Lemma 2.** If the $t$-th global iteration is a collaborative update we have

$$\mathbb{E}[||\bar{v}_t^{\psi(t)}||^2] = \mathbb{E}[||\bar{v}_t^{\psi(t)}||^2] = \mathbb{E}[||\bar{v}_t^{\psi(t)}||^2] = 2\mathbb{E}[||\bar{v}_t^{\psi(t)}||^2] + 2\mathbb{E}[||\bar{v}_t^{\psi(t)}||^2] - \mathbb{E}[||\bar{v}_t^{\psi(t)}||^2] + 2\mathbb{E}[||\bar{v}_t^{\psi(t)}||^2]$$

where (a) follows from $||a + b||^2 \leq 2||a||^2 + 2||b||^2$, (b) follows from Assumption 2, (c) follows from the Eq. 5, (d) follows from Assumption 3, (e) follows from definition of $L_2^2$, (f) follows from the definition of $\bar{v}_t$ and Assumption 4.

If the $t$-th global iteration is a dominated update, there is

$$\mathbb{E}[||\bar{v}_t^{\psi(t)}||^2] = \mathbb{E}[||\bar{v}_t^{\psi(t)}||^2] \leq G$$  \hspace{1cm} (23)
Then for \( \forall t \), according to Eqs. \ref{eq:22} and \ref{eq:23}, we have
\[
E[\|\tilde{\nu}^{(t)}_0\|^2] \leq G \leq 2G
\]
\[
E[\|\tilde{\nu}^{(1)}_t\|^2] \leq 2G + 2L^2\gamma^2 \tau_2 \sum_{t' \in D'(t)} E[\|\tilde{\nu}^{(t')}_{t'}\|^2] \leq 2G + 2L^2\gamma^2 \tau_2 E[\|\tilde{\nu}^{(0)}_0\|^2] \leq 2G \frac{1 - k^{1+1}}{1 - k}
\]
\[\ldots\]
\[
E[\|\tilde{\nu}^{(t)}_t\|^2] \leq 2G + 2L^2\gamma^2 \tau_2 \sum_{t' \in D'(t)} E[\|\tilde{\nu}^{(t')}_{t'}\|^2] \leq 2G + 2L^2\gamma^2 \tau_2 (2G \frac{1 - k^t}{1 - k}) \leq 2G \frac{1 - k^{t+1}}{1 - k}
\]
(24)
where (a) follows from that the 0-th global iteration must be a dominated update, (b) follows from that for all \( t' \in D'(t) \), there is \( t' \leq t \), (c) follows from that \( k := 2L^2\gamma^2 \tau_2 \), (d) follows from the summation formula of equal ratio sequence. According to Eq. \ref{eq:24} it holds that for \( \forall t \) there is
\[
E[\|\tilde{\nu}^{(t)}_t\|^2] \leq 2G \frac{1 - k^{t+1}}{1 - k} \leq \frac{2G}{1 - k} \leq \frac{2G}{1 - 2L^2\gamma^2 \tau}
\]
(25)
where the last inequality follows from the definition of \( \tau \). This completes the proof.

Lemma 3. For all \( \forall t \), there is
\[
E[\|\tilde{\nu}^{(t)}_t - \tilde{\nu}^{(t)}_t\|^2] \leq 2L^2\gamma^2 \tau_1 \sum_{t' \in D(t)} E[\|\tilde{\nu}^{(t')}_{t'}\|^2] + 8L^2\gamma^2 \tau_2 \sum_{t' \in D'(t)} E[\|\tilde{\nu}^{(t')}_{t'}\|^2]
\]
(26)

Proof of Lemma 3. First, we give the bound of \( E[\|\tilde{\nu}^{(t)}_t - \tilde{\nu}^{(t)}_t\|^2] \) as follow
\[
E[\|\tilde{\nu}^{(t)}_t - \tilde{\nu}^{(t)}_t\|^2] \leq E[\|\nabla_{\nu^{(t)}} f(\tilde{w}_t) - \nabla_{\nu^{(t)}} f(\tilde{w}_t) + \nabla_{\nu^{(t)}} g((\tilde{w}_t)\nu^{(t)}) - \nabla_{\nu^{(t)}} g((\tilde{w}_t)\nu^{(t)})\|^2
\]
\[
\leq 2E[\|\nabla_{\nu^{(t)}} f_i((\tilde{w}_t)\nu^{(t)}) - \nabla_{\nu^{(t)}} f_i(\tilde{w}_t)\|^2 + 2E[\|\nabla_{\nu^{(t)}} g((\tilde{w}_t)\nu^{(t)}) - \nabla_{\nu^{(t)}} g((\tilde{w}_t)\nu^{(t)})\|^2
\]
\[
\leq 2L^2E[\|\tilde{w}_t - \tilde{w}_t\|^2 + 2L^2\|\tilde{w}_t\|\|\tilde{w}_t\| - \|\tilde{w}_t\|\|\nu^{(t)}\|^2
\]
\[
\leq 2L^2\gamma^2 \sum_{t' \in D'(t)} E[\|\tilde{\nu}^{(t')}_{t'}\|^2] + 2L^2\gamma^2 \sum_{t' \in D'(t), \nu^{(t')} = \nu^{(t)}} E[\|\tilde{\nu}^{(t')}_{t'}\|^2]
\]
\[
\leq 2L^2\gamma^2 \sum_{t' \in D'(t)} E[\|\tilde{\nu}^{(t')}_{t'}\|^2] + 2L^2\gamma^2 \sum_{t' \in D'(t)} E[\|\tilde{\nu}^{(t')}_{t'}\|^2]
\]
\[
\leq 4L^2\gamma^2 \sum_{t' \in D'(t)} E[\|\tilde{\nu}^{(t')}_{t'}\|^2]
\]
(27)
where (a) follows from the definition of \( \tilde{\nu}^{(t)}_t \) and the definitions of \( \tilde{\nu}^{(t)}_t \) for different types of the \( t \)-th global iteration (i.e., dominated or collaborative), (b) follows from \( \|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2 \), (c) follows from Assumptions 1 and 2, (d) follows from Eq. \ref{eq:4}, (e) follows from Assumption 3 and \( \|\sum_{i=1}^na_i\|^2 \leq n \sum_{i=1}^n\|a_i\|^2 \), (f) follows from the definition of \( L_n \). Then we consider the bound of \( E[\|\tilde{\nu}^{(t)}_t - \tilde{\nu}^{(t)}_t\|^2] \):
\[
E[\|\tilde{\nu}^{(t)}_t - \tilde{\nu}^{(t)}_t\|^2] = E[\|\tilde{\nu}^{(t)}_t - \tilde{\nu}^{(t)}_t - \tilde{\nu}^{(t)}_t + \tilde{\nu}^{(t)}_t\|^2
\]
\[
\leq 2E[\|\tilde{\nu}^{(t)}_t - \tilde{\nu}^{(t)}_t\|^2 + 2E[\|\tilde{\nu}^{(t)}_t - \tilde{\nu}^{(t)}_t\|^2
\]
\[
\leq 2E[\|\nabla_{\nu^{(t)}} f_i(\tilde{w}_t) - \nabla_{\nu^{(t)}} f_i(\tilde{w}_t)\|^2 + 2E[\|\tilde{\nu}^{(t)}_t - \tilde{\nu}^{(t)}_t\|^2
\]
\[
\leq 2L^2E[\|\tilde{w}_t - \tilde{w}_t\|^2 + 2E[\|\tilde{\nu}^{(t)}_t - \tilde{\nu}^{(t)}_t\|^2
\]
\[
\leq 2L^2\gamma^2 \sum_{t' \in D'(t)} E[\|\tilde{\nu}^{(t')}_{t'}\|^2] + 2E[\|\tilde{\nu}^{(t)}_t - \tilde{\nu}^{(t)}_t\|^2
\]
\[
\leq 2L^2\gamma^2 \sum_{t' \in D'(t)} E[\|\tilde{\nu}^{(t')}_{t'}\|^2] + 2E[\|\tilde{\nu}^{(t)}_t - \tilde{\nu}^{(t)}_t\|^2
\]
\[
\leq 2L^2\gamma^2 \sum_{t' \in D'(t)} E[\|\tilde{\nu}^{(t')}_{t'}\|^2] + 2E[\|\tilde{\nu}^{(t)}_t - \tilde{\nu}^{(t)}_t\|^2
\]
\[
\leq 2L^2\gamma^2 \sum_{t' \in D'(t)} E[\|\tilde{\nu}^{(t')}_{t'}\|^2] + 2E[\|\tilde{\nu}^{(t)}_t - \tilde{\nu}^{(t)}_t\|^2
\]
\[
(\text{c)} \leq 2L^2\gamma^2\tau_1 \sum_{t' \in D(t)} \mathbb{E}\|\tilde{v}_u^{(t')}\|^2 + 8L^2\gamma^2\tau_2 \sum_{t' \in D'(t)} \mathbb{E}\|\tilde{v}_u^{(t')}\|^2 \tag{28}
\]

where (a) follows from \(\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2\), (b) follows from Assumptions 1, (c) follows from Eq. 4, (d) follows from Assumptions 3 and \(\sum_{i=1}^n a_i^2 \leq n \sum_{i=1}^n \|a_i\|^2\), (e) follows from the definition of \(L_a\) and Eq. 27. This completes the proof.

**Lemma 4.** For VFB\(^2\)-SGD, we have

\[
\sum_{u \in K(t)} \mathbb{E}\|\nabla g_{\psi(u)} f(w_t)\|^2 \geq \frac{1}{2} \mathbb{E}\|\nabla g_{\psi(u)} f(w_t)\|^2 - L^2\gamma^2\eta_1 \sum_{u \in K(t)} \sum_{u' \in \{t, \ldots, u\}} \mathbb{E}\|\tilde{v}_u^{(t')}\|^2
\]

**Proof of Lemma 5.** For any \(u \in K(t)\), there is

\[
\mathbb{E}\|\nabla g_{\psi(u)} f(w_t)\|^2 = \mathbb{E}\|\nabla g_{\psi(u)} f(w_t) - \nabla g_{\psi(u)} f(w_u) + \nabla g_{\psi(u)} f(w_u)\|^2
\]

\[
\overset{(a)}{\leq} 2\mathbb{E}\|\nabla g_{\psi(u)} f(w_t) - \nabla g_{\psi(u)} f(w_u)\|^2 + 2\mathbb{E}\|\nabla g_{\psi(u)} f(w_u)\|^2
\]

\[
\overset{(b)}{\leq} 2L^2\gamma^2\mathbb{E}\|w_t - w_u\|^2 + 2\mathbb{E}\|\nabla g_{\psi(u)} f(w_u)\|^2
\]

\[
\overset{(c)}{=} 2L^2\gamma^2\mathbb{E}\sum_{u \in \{t, \ldots, u\}} \mathbb{E}\|\tilde{v}_u^{(t')}\|^2 + 2\mathbb{E}\|\nabla g_{\psi(u)} f(w_u)\|^2
\]

\[
\overset{(d)}{\leq} 2L^2\gamma^2\eta_1 \sum_{u \in \{t, \ldots, u\}} \mathbb{E}\|\tilde{v}_u^{(t')}\|^2 + 2\mathbb{E}\|\nabla g_{\psi(u)} f(w_u)\|^2 \tag{29}
\]

where (a) follows from \(\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2\), (c) follows from Assumptions 1, (d) follows from Eq. 20, (e) follows from The bound of \(|K(t)|\) and \(\|\sum_{i=1}^n a_i\|^2 \leq n \sum_{i=1}^n \|a_i\|^2\). According to Eq. 29 we have

\[
\mathbb{E}\|\nabla g_{\psi(u)} f(w_t)\|^2 \geq \frac{1}{2} \mathbb{E}\|\nabla g_{\psi(u)} f(w_t)\|^2 - L^2\gamma^2\eta_1 \sum_{u' \in \{t, \ldots, u\}} \mathbb{E}\|\tilde{v}_{u'}^{(t')}\|^2 \tag{30}
\]

Summing above equality for all \(u \in K(t)\) we obtain the conclusion. This completes the proof.

**Proof of Theorem 7.** For \(\forall u \in K(t)\) we have that

\[
\mathbb{E}f(w_{u+1}) \tag{31}
\]

\[
\overset{(a)}{\leq} \mathbb{E} \left( f(w_u) + \langle \nabla f(w_u), w_{u+1} - w_u \rangle + \frac{L}{2} \|w_{u+1} - w_u\|^2 \right)
\]

\[
= \mathbb{E} \left( f(w_u) - \gamma \langle \nabla f(w_u), \tilde{v}_u^{(u)} \rangle \right) + \frac{L\gamma^2}{2} \|\tilde{v}_u^{(u)}\|^2
\]

\[
= \mathbb{E} \left( f(w_u) - \gamma \langle \nabla f(w_u), \tilde{v}_u^{(u)} + v_u^{(u)} - v_u^{(u)} \rangle \right) + \frac{L\gamma^2}{2} \|\tilde{v}_u^{(u)}\|^2
\]

\[
\overset{(b)}{=} \mathbb{E} f(w_u) - \gamma \mathbb{E}\langle \nabla f(w_u), \nabla g_{\psi(u)} f(w_u) \rangle + \frac{L\gamma^2}{2} \mathbb{E}\|\tilde{v}_u^{(u)}\|^2 + \gamma \mathbb{E}\|\nabla f(w_u)\|^2 + \mathbb{E}\|\tilde{v}_u^{(u)}\|^2 - \mathbb{E}\|\tilde{v}_u^{(u)}\|^2
\]

\[
\overset{(c)}{\leq} \mathbb{E} f(w_u) - \gamma \mathbb{E}\|\nabla g_{\psi(u)} f(w_u)\|^2 + \frac{\gamma^2}{2} \mathbb{E}\|\tilde{v}_u^{(u)}\|^2 + \frac{L\gamma^2}{2} \mathbb{E}\|\tilde{v}_u^{(u)}\|^2 + \gamma \mathbb{E}\|\nabla f(w_u)\|^2 + \mathbb{E}\|\tilde{v}_u^{(u)}\|^2 - \mathbb{E}\|\tilde{v}_u^{(u)}\|^2
\]

\[
\overset{(d)}{\leq} \mathbb{E} f(w_u) - \frac{\gamma}{2} \mathbb{E}\|\nabla g_{\psi(u)} f(w_u)\|^2 + \frac{L\gamma^2}{2} \mathbb{E}\|\tilde{v}_u^{(u)}\|^2 + L^2\gamma^2 \mathbb{E}\|\tilde{v}_u^{(u)}\|^2
\]

\[
+ L^2\gamma^2 \mathbb{E}\|\tilde{v}_u^{(u)}\|^2 + 4L^2\gamma^2\tau_2 \sum_{t' \in D'(t)} \mathbb{E}\|\tilde{v}_u^{(t')}\|^2
\]

where the inequalities (a) follows form Assumption 2, (b) follows from that \(v_u^{(u)} = \nabla g_{\psi(u)} f_{iu}(w_u)\) for a specific party, (c) follows from \(\langle a, b \rangle \leq \frac{1}{2}(\|a\|^2 + \|b\|^2)\), (d) follows from Lemma 3 and the definition of \(L_a\). Summing Eq. 31 over all \(u \in K(t)\),
we obtain
\[
\mathbb{E} [ f(w_{t+1}|K(t)) - f(w_t) ] \leq -\frac{\gamma}{2} \sum_{u \in K(t)} \mathbb{E} [\nabla g_{w(u)}^T \nabla f(w_u)]^2 + \frac{L_\ast \gamma^2}{2} \sum_{u \in K(t)} \mathbb{E} [\nabla \psi_u^\dagger]^2 \\
+ (L_\ast^2 \gamma^3 \tau_1 + 4L_\ast^2 \gamma^3 \tau_2) \sum_{u \in K(t), w' \in D'(u)} \mathbb{E} [\nabla \psi_{u'}^\dagger]^2
\]

where (a) follows from Lemma [4] (b) follows from Assumption [4]. According to Eq. [32] we have
\[
\mathbb{E} [ f(w_{t+1}|K(t)) - f(w^*) ] \leq (1 - \frac{\gamma \mu}{2}) (f(w_t) - f(w^*)) + C
\]

Assuming that \( U_{k \in P(t)} = \{0, 1, \ldots, t\} \), applying Eq. [33] we have that
\[
\mathbb{E} [ f(w_t) - f(w^*) ] \leq (1 - \frac{\gamma \mu}{2}) (f(w_0) - f(w^*)) + C \sum_{i=0}^{v(t)} (1 - \frac{\gamma \mu}{2})^i \\
\leq (1 - \frac{\gamma \mu}{2})^{v(t)} (f(w_0) - f(w^*)) + C \frac{(1 - (1 - \frac{\gamma \mu}{2})^{v(t)})}{\gamma \mu} \\
\leq (1 - \frac{\gamma \mu}{2})^{v(t)} (f(w_0) - f(w^*)) + C \frac{(1 - (1 - \frac{\gamma \mu}{2})^{v(t)})}{\gamma \mu} \\
\leq (1 - \frac{\gamma \mu}{2})^{v(t)} (f(w_0) - f(w^*)) + \left( \frac{L_\ast^2 \gamma^3 \tau_1}{2} + \frac{L_\ast \gamma^2 \tau^{1/2}}{2} + 5L_\ast^2 \gamma^3 \tau_3/2 \right) \frac{2G}{1 - 2L_\ast^2 \gamma^2 \tau} \gamma \mu,
\]

where (a) follows from the definition of \( C \). To obtain the \( \epsilon \) solution one can choose suitable \( \gamma \), such that
\[
1 - 2L_\ast^2 \gamma^2 \tau > 0
\]
\[
(1 - \frac{\gamma \mu}{2})^{v(t)} (f(w_0) - f(w^*)) \leq \frac{\epsilon}{2}
\]
\[
\left( \frac{L_\ast^2 \gamma^3 \tau_3/2}{2} + \frac{L_\ast \gamma^2 \tau^{1/2}}{2} + 5L_\ast^2 \gamma^3 \tau_3/2 \right) \frac{2G}{1 - 2L_\ast^2 \gamma^2 \tau} \gamma \mu \leq \frac{\epsilon}{2}.
\]

According to Eq. [35] there is \( \gamma^2 \leq \frac{1}{2L_\ast^2 \tau} \), which implies that \( \frac{L_\ast^2 \gamma^3 \tau_3/2}{2} + \frac{L_\ast \gamma^2 \tau^{1/2}}{2} + 5L_\ast^2 \gamma^3 \tau_3/2 \leq 6L_\ast^2 \gamma^3 \tau_3/2 \) (here we assume that \( L_\ast \) can be chosen a value \( \geq 1 \), this is reasonable from the definition of \( L_\ast \)). Thus, we can rewrite Eq. [37] as
\[
6L_\ast^2 \gamma^3 \tau_3/2 \frac{4G}{\mu(1 - 2L_\ast^2 \gamma^2 \tau)} \leq \frac{\epsilon}{2}.
\]
which implies that if $\tau$ is upper bounded, i.e., $\tau \leq \min\{e^{-4/3}, (GL^2)^{2/3}\}$, we can carefully choose $\gamma \leq \frac{e^{4/3}}{(GL^2)^{-1/3}}$ such that Eq. (37) holds. According to Eq. (38), there is

$$\log\left(\frac{2(f(w_0) - f(w^*))}{\epsilon}\right) \leq v(t)\log\left(\frac{1}{1 - 2\epsilon}\right)$$

(39)

Because $\log\left(\frac{1}{x}\right) \geq 1 - x$ for $0 < x \leq 1$, we have

$$v(t) \geq \frac{2}{\gamma\mu} \log\left(\frac{2(f(w_0) - f(w^*))}{\epsilon}\right) \geq \frac{44(2GL^2)^{1/3}}{\mu^4/3\epsilon} \log\left(\frac{2(f(w_0) - f(w^*))}{\epsilon}\right)$$

(40)

This completes the proof.

\[\square\]

**Proof of Theorem 2**

**Lemma 5.** For VFB$^2$-SVRG, let $u \in K(t)$ for all, we have that one can get:

$$\mathbb{E}\left\|\hat{\nu}_u^{\psi}(u)\right\|^2 \leq \frac{18G}{1 - 2L^2\gamma^2\tau^2}$$

(41)

**Proof of Lemma 5.** First, we prove the relation between $\mathbb{E}\left\|\hat{\nu}_u^{\psi}(u)\right\|^2$ and $\mathbb{E}\left\|\tilde{\nu}_u^{\psi}(u)\right\|^2$.

$$\mathbb{E}\left\|\hat{\nu}_u^{\psi}(u)\right\|^2 = \mathbb{E}\left\|\tilde{\nu}_u^{\psi}(u)\right\|^2 + \mathbb{E}\left\|\hat{\nu}_u^{\psi}(u) - \tilde{\nu}_u^{\psi}(u)\right\|^2$$

(42)

where (a) follows from $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. The upper bound to $\mathbb{E}\left\|\hat{\nu}_u^{\psi}(u) - \tilde{\nu}_u^{\psi}(u)\right\|^2$ can be obtained as follows.

$$\mathbb{E}\left\|\hat{\nu}_u^{\psi}(u) - \tilde{\nu}_u^{\psi}(u)\right\|^2 = \mathbb{E}\left\|\nabla G_t \hat{f}_t(\tilde{w}_u) - \nabla G_t \hat{f}_t(\tilde{w}_u)\right\|^2$$

(43)

where (a) follows from Assumption 2, (b) follows from Eq. (4), (c) follows from Assumption 3. Combining Eqs. (42) and (43), we have that

$$\mathbb{E}\left\|\tilde{\nu}_u^{\psi}(u)\right\|^2 \leq \frac{18G}{1 - 2L^2\gamma^2\tau^2}$$

(44)

Then following the analyses of follows from Lemma 2, we have

$$\mathbb{E}\left\|\tilde{\nu}_u^{\psi}(u)\right\|^2 \leq \frac{18G}{1 - 2L^2\gamma^2\tau^2}$$

This completes the proof.

\[\square\]

**Lemma 6.** Given the conditions in Theorem 2, let $u \in K(t)$, we have that:

$$\mathbb{E}\left\|\tilde{\nu}_u^{\psi}(u)\right\|^2 \leq \frac{16L^2}{\mu} \mathbb{E}\left(f(w^2) - f(w^*)\right) + \frac{8L^2}{\mu} \mathbb{E}\left(f(w^2) - f(w^*)\right) + 8L^2\gamma^2\eta_1 \sum_{u' \in \{t, ..., u\}} \mathbb{E}\left\|\tilde{\nu}_{u'}^{\psi}(u')\right\|^2$$

$$+ 4L^2\gamma^2\tau_1 \sum_{u' \in D(u)} \mathbb{E}\left\|\tilde{\nu}_{u'}^{\psi}(u')\right\|^2 + 16L^2\gamma^2\tau_2 \sum_{u' \in D^2(u)} \mathbb{E}\left\|\tilde{\nu}_{u'}^{\psi}(u')\right\|^2$$

(46)
Proof of Lemma 7. Define $E\|v_u^{(t)}\|^2 = E \|\nabla \varphi_{u,t} f_i (w_u^*) - \nabla \varphi_{w,t} f_i (w^*) + \nabla \varphi_{w,t} f (w^*)\|^2$, we have that $E\|v_u^{(t)}\|^2 = E\|v_u^{(t)} - v_u^{(t)}(u)\|^2 + 2E\|v_u^{(t)}\|^2$. First we give the upper bound to $E\|v_u^{(t)}\|^2$ as follows.

$$
E\|v_u^{(t)}\|^2 = E \left\| \nabla \varphi_{u,t} f_i (w_u^*) - \nabla \varphi_{w,t} f_i (w^*) + \nabla \varphi_{w,t} f (w^*) \right\|^2 \\
= E \left\| \nabla \varphi_{u,t} f_i (w_u^*) - \nabla \varphi_{w,t} f_i (w^*) + \nabla \varphi_{w,t} f (w^*) \right\|^2 \\
\leq 2E \left\| \nabla \varphi_{u,t} f_i (w_u^*) - \nabla \varphi_{w,t} f_i (w^*) \right\|^2 + 2E \left\| \nabla \varphi_{u,t} f (w^*) - \nabla \varphi_{w,t} f (w^*) + \nabla \varphi_{w,t} f (w^*) \right\|^2 \\
\leq 2E \left\| \nabla \varphi_{u,t} f_i (w_u^*) - \nabla \varphi_{w,t} f_i (w^*) \right\|^2 + 2E \left\| \nabla \varphi_{u,t} f (w^*) - \nabla \varphi_{w,t} f (w^*) \right\|^2 \\
\leq 2L^2E \|w_u^* - w^*\|^2 + 2L^2E \|w^* - w^*\|^2 \\
= 2L^2E \|w_u^* - w^* + w^* - w^*\|^2 + 2L^2E \|w^* - w^*\|^2 \\
\leq 4L^2E \|w_u^* - w^*\|^2 + 4L^2E \|w_u^* - w^*\|^2 + 2L^2E \|w^* - w^*\|^2 \\
\leq 4L^2 \gamma^2 E \left\| \sum_{\{t \in \{t, \ldots, \} \}} u \nabla \varphi_{u,t} v_u^{(u')} \right\|^2 + 4L^2E \|w_u^* - w^*\|^2 + 2L^2E \|w^* - w^*\|^2 \\
\leq 8L^2 \gamma^2 E \left\| f (w_u^*) - f (w^*) \right\|^2 + 4L^2 \frac{4L^2}{\mu} E \left\| f (w^*) - f (w^*) \right\|^2 + 4L^2 \gamma^2 \eta_1 \sum_{u \in \{t \in \{t, \ldots, \} \}} E \left\| v_u^{(u')} \right\|^2, \\
(47)
$$

where (a) and (c) follow from $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, (b) follows from Assumption 4, (d) follows from Eq. 20 and (e) follows from Assumption 4. Next we give the upper bound of $E\|v_u^{(t)}(u) - v_u^{(t)}\|^2$. Following the proof of Lemma 3, we have

$$
E\|v_u^{(t)}(u) - v_u^{(t)}\|^2 \leq 2L^2 \gamma^2 \tau_1 \sum_{t' \in D(t)} E\|v_u^{(t')}\|^2 + 8L^2 \gamma^2 \tau_2 \sum_{t' \in D'(t)} E\|v_u^{(t')}\|^2 \\
(48)
$$

combining above two equalities, we have

$$
E\|v_u^{(t)}\|^2 \leq \frac{16L^2}{\mu} E \left\| f (w_u^*) - f (w^*) \right\|^2 + 8L^2 \frac{4L^2}{\mu} E \left\| f (w^*) - f (w^*) \right\|^2 + 8L^2 \gamma^2 \eta_1 \sum_{u \in \{t \in \{t, \ldots, \} \}} E \left\| v_u^{(u')} \right\|^2 \\
+ 4L^2 \gamma^2 \tau_1 \sum_{u \in D(t)} E\|v_u^{(u')}\|^2 + 16L^2 \gamma^2 \tau_2 \sum_{u \in D'(t)} E\|v_u^{(u')}\|^2 \\
(49)
$$

This completes the proof.

Proof of Theorem 2. Similar to Eq. 31 for $u \in K(t)$ at $s$-th outer loop, we have that

$$
E_f (w_u^{(t+1)}) \\
\leq E \left\{ f (w_u^*) + \nabla f (w_u^*) (w_u^* - w_u^*) + \frac{L}{2} \|w_u^* - w_u^*\|^2 \right\} \\
= E \left\{ f (w_u^*) - \nabla f (w_u^*) (w_u^* - w_u^*) + \frac{L}{2} \|w_u^* - w_u^*\|^2 \right\} \\
= E \left\{ f (w_u^*) - \nabla f (w_u^*) (w_u^* - w_u^*) + \frac{L}{2} \|w_u^* - w_u^*\|^2 \right\} \\
= E \left\{ f (w_u^*) - \nabla f (w_u^*) (w_u^* - w_u^*) + \frac{L}{2} \|w_u^* - w_u^*\|^2 \right\} \\
\leq E \left\{ f (w_u^*) - \nabla f (w_u^*) (w_u^*) + \frac{L}{2} \|w_u^* - w_u^*\|^2 \right\} \\
(50)
$$

(51)
Summing Eq. **(50)** over all $u \in K(t)$, we obtain

$$
\mathbb{E} \left[ f(w_{t+1}^{K(t)}) - f(w^*) \right] 
\leq \frac{\gamma}{2} \sum_{u \in K(t)} \mathbb{E} \| \nabla g_{u}(f(w_u^*)) \|^2 + \frac{L_u \gamma}{2} \sum_{u \in K(t)} \mathbb{E} \| \nabla g_{u}(f(w_u^*)) \|^2 + (L_u^2 \gamma^2 \tau_1 + 4L_u^2 \gamma^2 \tau_2) \sum_{t' \in D(t)} \mathbb{E} \| \nabla g_{u}(f(w_{t'}^*)) \|^2
$$

$$
\leq \frac{\gamma}{2} \left( \sum_{u \in K(t)} \mathbb{E} \| \nabla g_{u}(f(w_u^*)) \|^2 - L_u^2 \gamma^2 \eta_1 \sum_{u \in K(t)} \sum_{u' \in \{t, \ldots, u\}} \mathbb{E} \| \nabla g_{u}(f(w_{u'}^*)) \|^2 \right)
+ \frac{L_u^2 \gamma^2}{2} \sum_{u \in K(t)} \mathbb{E} \| \nabla g_{u}(f(w_u^*)) \|^2 + (L_u^2 \gamma^2 \tau_1 + 4L_u^2 \gamma^2 \tau_2) \sum_{u \in K(t)} \sum_{u' \in D(t)} \mathbb{E} \| \nabla g_{u}(f(w_{u'}^*)) \|^2
$$

$$
= \frac{\gamma}{4} \sum_{u \in K(t)} \mathbb{E} \| \nabla g_{u}(f(w_u^*)) \|^2 + \frac{L_u^2 \gamma^2 \eta_1}{2} \sum_{u \in K(t)} \sum_{u' \in \{t, \ldots, u\}} \mathbb{E} \| \nabla g_{u}(f(w_{u'}^*)) \|^2
+ \frac{L_u^2 \gamma^2 \tau_1 + L_u^2 \gamma^2 \tau_2}{2} \sum_{u \in K(t)} \sum_{u' \in D(t)} \mathbb{E} \| \nabla g_{u}(f(w_{u'}^*)) \|^2
$$

$$
\leq \frac{\gamma \mu}{2} (f(w^*_t) - f(w^*)) + (L_u^2 \gamma^2 \tau_1 + 4L_u^2 \gamma^2 \tau_2) \sum_{u \in K(t)} \sum_{u' \in D(t)} \mathbb{E} \| \nabla g_{u}(f(w_{u'}^*)) \|^2
+ \frac{L_u^2 \gamma^2 \tau_1 + L_u^2 \gamma^2 \tau_2}{2} \sum_{u \in K(t)} \sum_{u' \in D(t)} \mathbb{E} \| \nabla g_{u}(f(w_{u'}^*)) \|^2
+ \frac{8L^2}{\mu} \mathbb{E} \| f(w^*_t) - f(w^*) \|^2 + \frac{8L^2}{\mu} \mathbb{E} \| f(w^*_t) - f(w^*) \|^2
$$

where (a) follows from Lemma **4**, (b) follows from Lemma **6**.

Let $e_t^s = \mathbb{E} (f(w^*_t) - f(w^*))$ and $e^s = \mathbb{E} (f(w^*_t) - f(w^*))$, we have

$$
\mathbb{E} \left[ \sum_{u \in K(t)} \mathbb{E} \left[ \left\| \nabla g_{u}(f(w_u^*)) \right\|^2 \right] \right]
\leq \left( \frac{1}{2} - \frac{\gamma \mu}{\mu} \right) e_t^s + \frac{8L \eta_1}{\mu} e^s + 8CL^2 \gamma^2 \eta_1 \sum_{u \in K(t)} \sum_{u' \in \{t, \ldots, u\}} \mathbb{E} \| \nabla g_{u}(f(w_{u'}^*)) \|^2
+ (5L^2 \gamma^2 \tau_1 + 4CL^2 \gamma^2 \tau_2 + 16CL^2 \gamma^2 \tau_2) \sum_{u \in K(t)} \sum_{u' \in D(t)} \mathbb{E} \| \nabla g_{u}(f(w_{u'}^*)) \|^2
$$

$$
\leq \left( \frac{1}{2} - \frac{\gamma \mu}{\mu} \right) e_t^s + \frac{8L \eta_1}{\mu} e^s + (28CL \gamma^2 \tau_1 + 5L \gamma^3 \tau_2) \frac{18G}{1 - 2L^2 \gamma^2 \tau}
$$

We carefully choose $\gamma$ such that $\frac{\gamma \mu}{\mu} = \frac{16L^2 \eta_1 C}{\mu} \leq \rho > 0$. Assume that $\cup_{u \in K(t)} = \{0, 1, \ldots, t\}$, applying above, we have

$$
\mathbb{E} \left[ \sum_{u \in K(t)} \mathbb{E} \left[ \left\| \nabla g_{u}(f(w_u^*)) \right\|^2 \right] \right]
\leq \left( \frac{1}{2} - \frac{\gamma \mu}{\mu} \right) e_t^s + \frac{8L \eta_1}{\mu} e^s + (28CL \gamma^2 \tau_1 + 5L \gamma^3 \tau_2) \frac{18G}{1 - 2L^2 \gamma^2 \tau}
+ \frac{8L \eta_1}{\mu} e^s + \frac{8L \eta_1}{\mu} e^s + (28CL \gamma^2 \tau_1 + 5L \gamma^3 \tau_2) \frac{18G}{1 - 2L^2 \gamma^2 \tau}
$$

$$
\leq \left( 1 - \rho \right) e^s + \frac{8L \eta_1}{\mu} e^s + (28CL \gamma^2 \tau_1 + 5L \gamma^3 \tau_2) \frac{18G}{1 - 2L^2 \gamma^2 \tau}
$$

Thus, to achieve the accuracy $\epsilon$, for VFB$_2$-SVRG, i.e., $\mathbb{E} f(w_S^t) - f(w^*_t) \leq \epsilon$, we can carefully choose $\gamma$ such that

$$
\frac{8L^2 \eta_1 C}{\rho \mu} \leq 0.05
\leq \frac{18G}{\rho (1 - 2L^2 \gamma^2 \tau)} \leq \frac{\epsilon}{8}
$$
And then let \((1 - \rho)v(t) \leq 0.25\), i.e., \(v(t) \geq \frac{\log 0.25}{\log(1 - \rho)}\), we have that 
\[
e^{s+1} \leq 0.75e^s + \frac{\epsilon}{8}
\]
Recursively apply above equality, we have that
\[
e^S \leq (0.75)^Se^0 + \frac{\epsilon}{2}
\]
Finally, the outer loop number \(S\) should satisfy the condition of \(S \geq \frac{\log 2\alpha_s}{\log \frac{\epsilon}{4}}\) and epoch number \(v(t)\) in an outer loop should satisfy \(v(t) \geq \frac{\log 0.25}{\log(1 - \rho)}\). This completes the proof. \(\square\)

**Proof of Theorem 3**

First we introduce following notations. \(\phi(t)\) denotes the corresponding local time counter on the party \(\psi(t)\). Given a local time counter \(u\) and \(\ell\)-th party, \(\xi(u, \ell)\) denotes the corresponding global time counter not only satisfying \(\phi(\xi(u, \ell)) = u\) but also \(\psi(\xi(u, \ell)) = \ell\).

**Lemma 7.** For \(\text{VFB}^2\text{-SAGA},\) we have that

\[
E\|\alpha_{t_i}^{t,\psi(t)} - \nabla g_{\psi(t)}f_{t_i}(w^*)\|^2 \\
\leq \frac{1}{n} \sum_{t' = 1}^{\phi(t) - 1} \sum_{i = 1}^{n} \frac{1}{n} \left(1 - \frac{1}{n}\right) \phi(t) - t' - 1 \|\nabla g_{\psi(t)}f_i(w_{\xi(t',\psi(t))}) - \nabla g_{\psi(t)}f_i(w^*)\|^2 \\
+ \frac{1}{n} \sum_{i = 1}^{n} \left(1 - \frac{1}{n}\right) \phi(t) - t' - 1 \|\nabla g_{\psi(t)}f_i(w_0) - \nabla g_{\psi(t)}f_i(w^*)\|^2 \tag{56}
\]

\[
E\|\alpha_{t_i}^{t,\psi(t)} - \nabla g_{\psi(t)}f_{t_i}(w^*)\|^2 \\
\leq \frac{1}{n} \sum_{t' = 1}^{\phi(t) - 1} \sum_{i = 1}^{n} \frac{1}{n} \left(1 - \frac{1}{n}\right) \phi(t) - t' - 1 \|\nabla g_{\psi(t)}f_i(w_{\xi(t',\psi(t))}) - \nabla g_{\psi(t)}f_i(w^*)\|^2 \\
+ \frac{1}{n} \sum_{i = 1}^{n} \left(1 - \frac{1}{n}\right) \phi(t) - t' - 1 \|\nabla g_{\psi(t)}f_i(w_0) - \nabla g_{\psi(t)}f_i(w^*)\|^2 \tag{57}
\]

\[
E\|\alpha_{t_i}^{t,\psi(t)} - \hat{\alpha}_{t_i}^{t,\psi(t)}\|^2 \\
\leq \frac{\gamma_1 L^2}{n} \sum_{t' = 1}^{\phi(t) - 1} \sum_{u \in D(\xi(t',\psi(t)))} \left(1 - \frac{1}{n}\right) \phi(t) - t' - 1 \|\hat{\psi}_u\|^2 \tag{58}
\]

\[
E\|\alpha_{t_i}^{t,\psi(t)} - \hat{\alpha}_{t_i}^{t,\psi(t)}\|^2 \\
\leq \frac{4\gamma_2 L^2}{n} \sum_{t' = 1}^{\phi(t) - 1} \sum_{u \in D'(\xi(t',\psi(t)))} \left(1 - \frac{1}{n}\right) \phi(t) - t' - 1 \|\hat{\psi}_u\|^2 \tag{59}
\]

**Proof of Lemma 7.** Firstly, we have that

\[
E\|\alpha_{t_i}^{t,\psi(t)} - \nabla g_{\psi(t)}f_{t_i}(w_t)\|^2 = \frac{1}{n} \sum_{i = 1}^{n} E\|\hat{\alpha}_{t_i}^{t,\psi(t)} - \nabla g_{\psi(t)}f_i(w_t)\|^2 \tag{60}
\]

\[
= \frac{1}{n} \sum_{i = 1}^{n} \sum_{t' = 0}^{\phi(t) - 1} 1_{\{u^*_t = t\}} \|\nabla g_{\psi(t)}f_i(w_{\xi(t',\psi(t))}) - \nabla g_{\psi(t)}f_i(w^*)\|^2 \\
= \frac{1}{n} \sum_{t' = 0}^{\phi(t) - 1} \sum_{i = 1}^{n} E1_{\{u^*_t = t\}} \|\nabla g_{\psi(t)}f_i(w_{\xi(t',\psi(t))}) - \nabla g_{\psi(t)}f_i(w^*)\|^2
\]

where \(u^*_t\) denote the last iterate to update the \(\hat{\alpha}_{t_i}^{t,\psi(t)}\). Note that, we do not distinguish \(\psi(t)\) and \(\psi(t')\) because they correspond to the same party. We consider two cases including \(t' > 0\) and \(t' = 0\) as follows.
For $t' > 0$, we have that
\[
\mathbb{E} \left( \mathbf{1}_{\{u^*_i = t'\}} \left\| \nabla_{\tilde{w}_i} f_i(\tilde{w}_{\xi(t')}) - \nabla_{\tilde{w}_i} f_i(w_{t', \psi(t)}) \right\|^2 \right) \leq \mathbb{E} \left( \mathbf{1}_{\{i', t'\}} \sum_{i=1}^n \mathbb{E} \left[ \left( 1 - \frac{1}{n} \right)^{\phi(t') - 1} \| \nabla_{\tilde{w}_i} f_i(\tilde{w}_{\xi(t')}) - \nabla_{\tilde{w}_i} f_i(w^*) \|^2 \right] \right)
\]
where the inequality (a) uses the fact $\tilde{w}$ and $\tilde{w}$ are independent for $\psi \neq t'$, the inequality (b) uses the fact that $P\{i_t = i\} = \frac{1}{n}$ and $P\{i_t \neq i\} = 1 - \frac{1}{n}$.

For $t' = 0$, we have that
\[
\mathbb{E} \left( \mathbf{1}_{\{u^*_i = 0\}} \left\| \nabla_{\tilde{w}_i} f_i(\tilde{w}_0) - \nabla_{\tilde{w}_i} f_i(w_{t, \psi(t)}) \right\|^2 \right) \leq \mathbb{E} \left( \mathbf{1}_{\{i_t \neq i\}} \sum_{i=1}^n \mathbb{E} \left[ \left( 1 - \frac{1}{n} \right)^{\phi(t) - 1} \| \nabla_{\tilde{w}_i} f_i(\tilde{w}_{\xi(t')} - \nabla_{\tilde{w}_i} f_i(w^*) \|^2 \right] \right)
\]

Substituting Eqs. (62) and (61) into (60), we have:
\[
\mathbb{E} \left[ \left\| \alpha_{t', \psi(t)} - \nabla_{\tilde{w}_{\xi(t')}} f_i(w^*) \right\|^2 \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[ \sum_{t'=0}^{\phi(t) - 1} \left( 1 - \frac{1}{n} \right)^{\phi(t) - 1} \mathbb{E} \left[ \left\| \nabla_{\tilde{w}_i} f_i(\tilde{w}_{\xi(t')} - \nabla_{\tilde{w}_i} f_i(\tilde{w}_0) - \nabla_{\tilde{w}_i} f_i(w^*) \|^2 \right] \right] \right]
\]
Combing the fact that $w$ represents the consistent read and thus $\tau_3 = 0$ with (63), we have
\[
\mathbb{E} \left[ \left\| \alpha_{t', \psi(t)} - \nabla_{\tilde{w}_{\xi(t')}} f_i(w^*) \right\|^2 \right] \leq \frac{L^2}{n} \sum_{t'=0}^{\phi(t) - 1} \left( 1 - \frac{1}{n} \right)^{\phi(t) - 1} \sigma(w_{\xi(t')} + L^2 \left( 1 - \frac{1}{n} \right)^{\phi(t)} \sigma(w_0),
\]
where $\sigma(w_u) = \mathbb{E} \| w_u - w^* \|^2$. Similarly, we have that
\[
\mathbb{E} \left[ \left\| \alpha_{t', \psi(t)} - \tilde{\alpha}_{t', \psi(t)} \right\|^2 \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[ \left\| \alpha_{t', \psi(t)} - \tilde{\alpha}_{t', \psi(t)} \right\|^2 \right]
\]
\[
= \frac{1}{n} \sum_{i=1}^n \sum_{t'=0}^{\phi(t) - 1} \mathbb{E} \left[ \left\| \alpha_{t', \psi(t')} - \tilde{\alpha}_{t', \psi(t')} \right\|^2 \right]
\]
where the inequality (a) can be obtained similar to (63) (note that \( \hat{a} \) inequality (d) uses Assumption 3. Moreover, we have

\[
\begin{align*}
\sum_{i=1}^{n} \sum_{t'=1}^{n} \left( \frac{1}{n} \right) \frac{\phi(t) - \phi(t')}{n} \mathbb{E} \left\| \alpha_i^{t', \psi(t')} - \hat{\alpha}_i^{t', \psi(t')} \right\|^2 \\
+ \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{n} \right) \frac{\phi(t)}{n} \mathbb{E} \left\| \alpha_i^{0, \psi} - \hat{\alpha}_i^{0, \psi} \right\|^2 \\
\end{align*}
\]

\[
\begin{align*}
\sum_{i=1}^{n} \sum_{t'=1}^{n} \left( \frac{1}{n} \right) \frac{\phi(t) - \phi(t')}{n} \mathbb{E} \left\| \alpha_i^{t', \psi(t')} - \hat{\alpha}_i^{t', \psi(t')} \right\|^2 \\
= \frac{1}{n} \sum_{i=1}^{n} \sum_{t'=1}^{n} \left( \frac{1}{n} \right) \frac{\phi(t) - \phi(t')}{n} \mathbb{E} \left\| \nabla f_i \left( \tilde{w}_{\xi(t', \psi(t))} \right) - \nabla f_i \left( \tilde{w}_{\xi(t', \psi(t))} \right) \right\|^2 \\
\leq \frac{L^2}{n} \sum_{i=1}^{n} \sum_{t'=1}^{n} \left( \frac{1}{n} \right) \frac{\phi(t) - \phi(t')}{n} \mathbb{E} \left\| \sum_{u \in D(\xi(t', \psi(t)))} U_{\psi(u)} \bar{v}_{\psi(u)} \right\|^2 \\
\leq \frac{\tau L^2}{n} \frac{\phi(t) - \phi(t')}{n} \mathbb{E} \left\| \nabla f_i \left( \tilde{w}_{\xi(t', \psi(t))} \right) - \nabla f_i \left( \tilde{w}_{\xi(t', \psi(t))} \right) \right\|^2 \\
\leq \frac{L^2}{n} \sum_{i=1}^{n} \sum_{t'=1}^{n} \left( \frac{1}{n} \right) \frac{\phi(t) - \phi(t')}{n} \mathbb{E} \left\| \alpha_i^{t', \psi(t')} - \hat{\alpha}_i^{t', \psi(t')} \right\|^2 \\
= \frac{1}{n} \sum_{i=1}^{n} \sum_{t'=1}^{n} \left( \frac{1}{n} \right) \frac{\phi(t) - \phi(t')}{n} \mathbb{E} \left\| \alpha_i^{t', \psi(t')} - \hat{\alpha}_i^{t', \psi(t')} \right\|^2 \\
\end{align*}
\]

(65)

\[
\begin{align*}
\sum_{i=1}^{n} \sum_{t'=1}^{n} \left( \frac{1}{n} \right) \frac{\phi(t) - \phi(t')}{n} \mathbb{E} \left\| \alpha_i^{t', \psi(t')} - \hat{\alpha}_i^{t', \psi(t')} \right\|^2 \\
= \frac{1}{n} \sum_{i=1}^{n} \sum_{t'=1}^{n} \left( \frac{1}{n} \right) \frac{\phi(t) - \phi(t')}{n} \mathbb{E} \left\| \alpha_i^{t', \psi(t')} - \hat{\alpha}_i^{t', \psi(t')} \right\|^2 \\
\end{align*}
\]

(65)

\[
\begin{align*}
\sum_{i=1}^{n} \sum_{t'=1}^{n} \left( \frac{1}{n} \right) \frac{\phi(t) - \phi(t')}{n} \mathbb{E} \left\| \alpha_i^{t', \psi(t')} - \hat{\alpha}_i^{t', \psi(t')} \right\|^2 \\
= \frac{1}{n} \sum_{i=1}^{n} \sum_{t'=1}^{n} \left( \frac{1}{n} \right) \frac{\phi(t) - \phi(t')}{n} \mathbb{E} \left\| \alpha_i^{t', \psi(t')} - \hat{\alpha}_i^{t', \psi(t')} \right\|^2 \\
\end{align*}
\]

(65)
(a)-(d) can be obtained from the analyses of Lemma 3. This completes the proof.

Lemma 8. Given a global iteration number $u$, we let \{ $\bar{u}_0, \bar{u}_1, \ldots, \bar{u}_{v(u)-1}$ \} be the all start iteration number for the global time counters from $0$ to $u$. Thus, for VFB$^2$-SAGA, we have that

$$E\|v^{(u)}\|^2 \leq \frac{4L^2}{n} \sum_{i=1}^{n} \sum_{u'=1}^{v(u)-1} \left(1 - \frac{1}{n}\right)^{\phi(u) - u'} \sigma(w_{u',v}) + 2L^2 \left(1 - \frac{1}{n}\right)^{\phi(u)} \sigma(w_0) + 4L^2 \sigma(w_{v(u)}) + 8L^2 \gamma^2 \eta_t^2 q G \tag{66}$$

Proof of Lemma 8. We have that

$$E\|v^{(u)}\|^2 = E\left\| \nabla_{\psi(u)} f_i(w^*) - \alpha_i^{(u)} + \frac{1}{n} \sum_{i=1}^{n} \alpha_i^{(u)} \right\|^2$$

$$= E\left\| \nabla_{\psi(u)} f_i(w_u) - \nabla_{\psi(u)} f_i(w^*) - \alpha_i^{u' \ell} + \frac{1}{n} \sum_{i=1}^{n} \alpha_i^{u' \ell} - \nabla_{\psi(u)} f_i(w_u) - \nabla_{\psi(u)} f_i(w^*) \right\|^2$$

$$\leq 2E\left\| \nabla_{\psi(u)} f_i(w_u) - \alpha_i^{u' \ell} + \frac{1}{n} \sum_{i=1}^{n} \alpha_i^{u' \ell} - \nabla_{\psi(u)} f_i(w_u) - \nabla_{\psi(u)} f_i(w^*) \right\|^2 + 2E\left\| \nabla_{\psi(u)} f_i(w_u) - \nabla_{\psi(u)} f_i(w^*) \right\|^2$$

$$\leq 2L^2 \left(1 - \frac{1}{n}\right)^{\phi(u) - u' - 1} E\left\| w_{\ell, u'} - w^* \right\|^2 + 2L^2 \left(1 - \frac{1}{n}\right)^{\phi(u)} \sigma(w_0)$$

$$+ 2L^2 E\|w_u - w^*\|^2$$
\[ + 2L^2 \left(1 - \frac{1}{n}\right)^{v(u)} \sigma(w_0) + 4L^2E\|w_{\varphi(u)} - w^*\|^2 + 8L^2\gamma^2\eta^2_{\lambda_2}G \]
\[ \leq 4\frac{L^2\eta}{n} \sum_{k'=1}^{v(u)} \left(1 - \frac{1}{n}\right)^{v(u)-k'} \sigma \left( w_{u_{k'}} \right) + 2L^2 \left(1 - \frac{1}{n}\right)^{v(u)} \sigma \left( w_0 \right) + 4L^2\gamma \sigma \left( w_{\varphi(u)} \right) + 8L^2\gamma^2\eta^2_{\lambda_2}G \]

where (a) and (d) uses \( \|\sum_{i=1}^{n}a_i^2\|^2 \leq n \sum_{i=1}^{n} \|a_i\|^2 \), (b) follows from \( E\|x - Ex\|^2 \leq E\|x\|^2 \), (c) uses Lemma [7], (e) follows from the bound of \( |K(t)| \), and (f) follows from Assumption [1]. (g) uses the fact \( \sum_{u=1}^{n} \frac{\phi(u)-u-1}{n} \leq n \).

Lemma 9. For all \( \forall \psi(t) \), there are

\[ E\|\tilde{\nu}_t^{\psi(t)}\|^2 \leq \lambda_7 G \]

where \( \lambda_7 = \frac{18}{1 - 2L_2\gamma^2\tau^2} \)

Proof of Lemma [9], we give the upper bound to \( E\|\tilde{\nu}_t^{\psi(u)} - \tilde{\alpha}_t^{\psi(u)}\|^2 \) as follows. We have that

\[ E\|\tilde{\nu}_t^{\psi(u)} - \tilde{\alpha}_t^{\psi(u)}\|^2 \]
\[ = E \left\| \left( \nabla \varphi_{\psi(t)} \mathcal{L}(\tilde{w}) + \nabla \varphi_{\psi(t)} g(\tilde{w}_t \nabla \varphi_{\psi(t)}) - \nabla \varphi_{\psi(t)} f(\tilde{w}_t) - \tilde{\alpha}_t^{\psi(u)} + \tilde{\alpha}_t^{\psi(u)} \right) + \frac{1}{n} \sum_{i=1}^{n} \tilde{\alpha}_i^{\psi(u)} - \frac{1}{n} \sum_{i=1}^{n} \tilde{\alpha}_i^{\psi(u)} \right\|^2 \]
\[ \leq 3E Q_1 + 3E \left\| \tilde{\alpha}_t^{\psi(u)} + \tilde{\alpha}_t^{\psi(u)} \right\|^2 + 3E \left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{\alpha}_i^{\psi(u)} - \frac{1}{n} \sum_{i=1}^{n} \tilde{\alpha}_i^{\psi(u)} \right\|^2 

where \( Q_1 = \left\| \left( \nabla \varphi_{\psi(t)} \mathcal{L}(\tilde{w}) + \nabla \varphi_{\psi(t)} g(\tilde{w}_t \nabla \varphi_{\psi(t)}) - \nabla \varphi_{\psi(t)} f(\tilde{w}_t) \right) + \frac{1}{n} \sum_{i=1}^{n} \tilde{\alpha}_i^{\psi(u)} - \frac{1}{n} \sum_{i=1}^{n} \tilde{\alpha}_i^{\psi(u)} \right\|^2 \) and inequality (a) uses \( \sum_{i=1}^{n} \alpha_i \leq n \sum_{i=1}^{n} \|a_i\|^2 \). We will give the upper bounds for the expectations of \( Q_1, Q_2 \) and \( Q_3 \) respectively.

\[ EQ_1 = E \left\| \left( \nabla \varphi_{\psi(t)} \mathcal{L}(\tilde{w}) + \nabla \varphi_{\psi(t)} g(\tilde{w}_t \nabla \varphi_{\psi(t)}) - \nabla \varphi_{\psi(t)} f(\tilde{w}_t) \right) + \frac{1}{n} \sum_{i=1}^{n} \tilde{\alpha}_i^{\psi(u)} - \frac{1}{n} \sum_{i=1}^{n} \tilde{\alpha}_i^{\psi(u)} \right\|^2 \]
\[ \leq (a) E \|\nabla \varphi_{\psi(t)} f(\tilde{w}_t) - \nabla \varphi_{\psi(t)} f(\tilde{w}_t) + \nabla \varphi_{\psi(t)} g(\tilde{w}_t \nabla \varphi_{\psi(t)}) - \nabla \varphi_{\psi(t)} g(\tilde{w}_t \nabla \varphi_{\psi(t)}) \|^2 \]
\[ \leq (b) 4L^2\gamma^2\tau^2 \sum_{t' \in D'(t)} E\|\tilde{\nu}_t^{\psi(t')}\|^2 \]

above inequality can be obtained by following the proof of Lemma [3]

\[ EQ_2 = E \left\| \tilde{\alpha}_t^{\psi(t)} - \tilde{\alpha}_t^{\psi(t')} \right\|^2 \]
\[ \leq \frac{4\tau^2 L^2\gamma^2}{n} \sum_{t' = 1}^{\phi(t)-1} \sum_{u \in D'(\xi(t',\psi(t)))} \left(1 - \frac{1}{n}\right)^{\phi(t')-1} E\|\tilde{\nu}_u^{\psi}\|^2 \]

where the inequality uses Lemma [7]

\[ EQ_3 = E \left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{\alpha}_i^{\psi(t)} - \frac{1}{n} \sum_{i=1}^{n} \tilde{\alpha}_i^{\psi(t)} \right\|^2 \]
\[ \leq \frac{1}{n} \sum_{i=1}^{n} E\|\tilde{\alpha}_i^{\psi(t)} - \tilde{\alpha}_i^{\psi(t)}\|^2 \]
\[ \leq \frac{4\tau^2 L^2\gamma^2}{n} \sum_{t' = 1}^{\phi(t)-1} \sum_{u \in D'(\xi(t',\psi(t)))} \left(1 - \frac{1}{n}\right)^{\phi(t')-1} E\|\tilde{\nu}_u^{\psi}\|^2 \]
where the first inequality uses \( \| \sum_{i=1}^{n} a_i \| \leq n \sum_{i=1}^{n} \| a_i \| \), the second inequality uses Lemma \textcolor{red}{7}. Combining \textcolor{red}{70, 71} and \textcolor{red}{72} one can obtain:

\[
\begin{align*}
\mathbb{E} \left\| \tilde{v}_t^{(t)} - \tilde{v}_t^{(t)} \right\|^2 & \leq 3 \mathbb{E} Q_1 + 3 \mathbb{E} Q_2 + 3 \mathbb{E} Q_3 \\
& \leq 12 L^2_{2} \gamma^2 \tau_2 \sum_{u \in D'}(\xi(t', \psi(t))) \mathbb{E} \| \tilde{v}_u^{(u)} \| ^2 + \frac{24r_2 L^4_{2} \gamma^2}{n} \sum_{t'=1}^{\phi(t)-1} \sum_{u \in D' (\xi(t', \psi(t)))} \left( 1 - \frac{1}{n} \right)^{\phi(t') - t' - 1} \mathbb{E} \| \tilde{v}_u^{(u)} \| ^2
\end{align*}
\]

Combining \( \mathbb{E} \| \tilde{v}_t^{(t)} \|^2 \leq 2 \mathbb{E} \| \tilde{v}_t^{(t)} - \tilde{v}_t^{(t)} \|^2 + 2 \mathbb{E} \| \tilde{v}_t^{(t)} \|^2 \) with Eq. \textcolor{red}{73} and following the analyses of Lemma \textcolor{red}{5} we have

\[
\mathbb{E} \| \tilde{v}_t^{(t)} \|^2 \leq \frac{18G}{1 - 72 L^2_{2} \gamma^2 \tau^2},
\]

This completes the proof. \( \square \)

Moreover, define \( v_t^{(t)} = \nabla_{\psi(t)} f_t(w^*) - \alpha_t^{(t)} + \frac{1}{n} \sum_{i=1}^{n} \alpha_i^{(t)} \). And then, we give the upper bound to \( \mathbb{E} \left\| \tilde{v}_t^{(t)} - v_t^{(t)} \right\|^2 \) as follows. We have that

\[
\mathbb{E} \left\| \tilde{v}_t^{(t)} - v_t^{(t)} \right\|^2 = \mathbb{E} \left\| \nabla_{\psi(t)} f_t(\tilde{w}_t) - \nabla_{\psi(t)} f_t(w^*) - \frac{1}{n} \sum_{i=1}^{n} \alpha_i^{(t)} - \nabla_{\psi(t)} f_t(\tilde{w}_t) + \frac{1}{n} \sum_{i=1}^{n} \alpha_i^{(t)} \right\|^2
\]

\[
\leq 3 \mathbb{E} \left\| \nabla_{\psi(t)} f_t(\tilde{w}_t) - \nabla_{\psi(t)} f_t(\tilde{w}_t) \right\|^2 + 3 \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \alpha_i^{(t)} \right\|^2
\]

where the inequality (a) uses \( \| \sum_{i=1}^{n} a_i \|^2 \leq n \sum_{i=1}^{n} \| a_i \|^2 \). We will give the upper bounds for the expectations of \( Q_4, Q_5 \) and \( Q_6 \) respectively.

\[
Q_4 = \mathbb{E} \left\| \nabla_{\psi(t)} f_t(\tilde{w}_t) - \nabla_{\psi(t)} f_t(w^*) \right\|^2 \leq L^2_{2} \gamma \mathbb{E} \| w^* - \tilde{w}_t \|^2 = L^2_{2} \gamma \mathbb{E} \left\| \sum_{t' \in D(\tilde{u})} U_{\psi(t')} \tilde{v}_t^{(t')} \right\|^2 \leq \tau t L^2_{2} \gamma \sum_{t' \in D(\tilde{u})} \mathbb{E} \left\| \tilde{v}_t^{(t')} \right\|^2
\]

where (a) uses Assumption 2, (b) uses \( \| \sum_{i=1}^{n} a_i \|^2 \leq n \sum_{i=1}^{n} \| a_i \|^2 \). Similar to the analyses of \( Q_2 \) and \( Q_3 \), we have

\[
Q_5 = \mathbb{E} \left\| \alpha_t^{(t)} - \tilde{\alpha}_t^{(t)} \right\|^2 \leq \frac{\tau t L^2_{2} \gamma^2}{n} \sum_{t'=1}^{\phi(t)-1} \sum_{\tilde{u} \in D(\xi(t', \psi(t'')))} \left( 1 - \frac{1}{n} \right)^{\phi(t') - t' - 1} \mathbb{E} \left\| \tilde{v}_{\tilde{u}}^{(\tilde{u})} \right\|^2
\]

where the inequality uses Lemma \textcolor{red}{7}.

\[
Q_6 = \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{\alpha}_i^{(t)} - \frac{1}{n} \sum_{i=1}^{n} \alpha_i^{(t)} \right\|^2 \leq \frac{\tau t L^2_{2} \gamma^2}{n} \sum_{t'=1}^{\phi(t)-1} \sum_{\tilde{u} \in D(\xi(t', \psi(t'')))} \left( 1 - \frac{1}{n} \right)^{\phi(t') - t' - 1} \mathbb{E} \left\| \tilde{v}_{\tilde{u}}^{(\tilde{u})} \right\|^2
\]
Based on above formulations, we have

\[
\mathbb{E} \left[ \left\| \bar{v}_t^{(\psi)} \right\|^2 \right] = \mathbb{E} \left[ \left\| \bar{v}_t^{(\psi)} - v_t^{(\psi)} + v_t^{(\psi)} \right\|^2 \right] \leq 2\mathbb{E} \left[ \left\| \bar{v}_t^{(\psi)} - v_t^{(\psi)} \right\|^2 \right] + 2\mathbb{E} \left[ \left\| v_t^{(\psi)} \right\|^2 \right] \\
\leq 2 \left( 2\mathbb{E} \left[ \left\| \bar{v}_u^{(\psi)} - v_u^{(\psi)} \right\|^2 \right] + 2\mathbb{E} \left[ \left\| v_u^{(\psi)} \right\|^2 \right] \right) + 2\mathbb{E} \left[ \left\| v_t^{(\psi)} \right\|^2 \right] \\
\leq \frac{360L^22\gamma^2\tau G}{1 - 72L^22\gamma^2\tau} + 2\mathbb{E} \left[ \left\| v_t^{(\psi)} \right\|^2 \right]
\]

(77)

**Proof of Theorem 3**

First, we upper bound \( \mathbb{E} f(w_{t+1}) \) for \( t = 0, \ldots, S - 1 \):

\[
\mathbb{E} \left[ f(w_{t+1}^K(t)) - f(w_t) \right]
\leq -\frac{\gamma}{2} \sum_{u \in K(t)} \mathbb{E} \left[ \left\| \nabla g_{\phi(u)}(w_u) \right\|^2 \right] + \frac{L^2\gamma^2}{2} \sum_{u \in K(t)} \mathbb{E} \left[ \left\| \bar{v}_u^{(\psi)} \right\|^2 \right] \\
+ \left( \frac{7L^2\gamma^3}{n} + \frac{4\gamma L^2\gamma^3}{n} \right) \sum_{u \in K(t)} \sum_{t' = 1}^{\phi(u)-1} \sum_{u' \in D'(\xi(t', \psi(t)))} \left( 1 - \frac{1}{n} \right)^{\phi(u) - t' - 1} \mathbb{E} \left[ \left\| \bar{v}_u^{(\psi)(\psi)} \right\|^2 \right]
\leq -\frac{\gamma}{2} \sum_{u \in K(t)} \mathbb{E} \left[ \left\| \nabla g_{\phi(u)}(w_u) \right\|^2 \right] - L^2\gamma^2 \eta_1 \sum_{u \in K(t)} \sum_{u' \in \{t, \ldots, n\}} \mathbb{E} \left[ \left\| \bar{v}_u^{(\psi)(\psi)} \right\|^2 \right] \\
+ \frac{L^2\gamma^2}{2} \sum_{u \in K(t)} \mathbb{E} \left[ \left\| \bar{v}_u^{(\psi)} \right\|^2 \right] + \frac{5\gamma L^2\gamma^3}{n} \sum_{u \in K(t)} \sum_{t' = 1}^{\phi(u)-1} \sum_{u' \in D'(\xi(t', \psi(t)))} \left( 1 - \frac{1}{n} \right)^{\phi(u) - t' - 1} \mathbb{E} \left[ \left\| \bar{v}_u^{(\psi)(\psi)} \right\|^2 \right]
\leq -\frac{\gamma}{4} \mathbb{E} \left[ \left\| \nabla f(w_t) \right\|^2 \right] + \left( \frac{L^2\gamma^3}{n} + \frac{L^2\gamma^2}{2} \right) \sum_{u \in K(t)} \mathbb{E} \left[ \left\| \bar{v}_u^{(\psi)} \right\|^2 \right] \\
+ \frac{5\gamma L^2\gamma^3}{n} \sum_{u \in K(t)} \sum_{t' = 1}^{\phi(u)-1} \sum_{u' \in D'(\xi(t', \psi(t)))} \left( 1 - \frac{1}{n} \right)^{\phi(u) - t' - 1} \mathbb{E} \left[ \left\| \bar{v}_u^{(\psi)(\psi)} \right\|^2 \right]
\leq -\frac{\gamma}{4} \mathbb{E} \left[ \left\| \nabla f(w_t) \right\|^2 \right] + \left( 2L^2\gamma^3 \frac{\tau}{1 \frac{18G}{1 - 72L^22\gamma^2\tau} + 2\mathbb{E} \left[ \left\| v_t^{(\psi)} \right\|^2 \right] \right) \\
+ \frac{L^2\gamma^3}{2} + \frac{L^2\gamma^2}{2} \sum_{u \in K(t)} \left( \frac{4L^2\gamma^4}{n} \sum_{u' = 1}^{\psi(u)-1} \sum_{u' = 1}^{\psi(u)-1} \left( 1 - \frac{1}{n} \right)^{\phi(u) - k' - 1} \mathbb{E} \left[ \left\| \bar{v}_u^{(\psi)} \right\|^2 \right]
\leq -\frac{\gamma}{4} \mathbb{E} \left[ \left\| \nabla f(w_t) \right\|^2 \right] + \left( 2L^2\gamma^3 \frac{\tau}{1 \frac{18G}{1 - 72L^22\gamma^2\tau} + 2\mathbb{E} \left[ \left\| v_t^{(\psi)} \right\|^2 \right] \right) \\
+ \frac{L^2\gamma^3}{2} + \frac{L^2\gamma^2}{2} \sum_{u \in K(t)} \left( \frac{4L^2\gamma^4}{n} \sum_{u' = 1}^{\psi(u)-1} \sum_{u' = 1}^{\psi(u)-1} \left( 1 - \frac{1}{n} \right)^{\phi(u) - k' - 1} \mathbb{E} \left[ \left\| \bar{v}_u^{(\psi)} \right\|^2 \right]
\leq -\frac{\gamma}{4} \mathbb{E} \left[ \left\| \nabla f(w_t) \right\|^2 \right] - \frac{\gamma}{2} \mathbb{E} \left[ \left\| \nabla g_{\phi(u)}(w_u) \right\|^2 \right] + \frac{L^2\gamma^3}{2} + \frac{L^2\gamma^2}{2} \sum_{u \in K(t)} \left( \frac{4L^2\gamma^4}{n} \sum_{u' = 1}^{\psi(u)-1} \sum_{u' = 1}^{\psi(u)-1} \left( 1 - \frac{1}{n} \right)^{\phi(u) - k' - 1} \mathbb{E} \left[ \left\| \bar{v}_u^{(\psi)} \right\|^2 \right]
\leq -\frac{\gamma}{4} \mathbb{E} \left[ \left\| \nabla f(w_t) \right\|^2 \right] + \left( 2L^2\gamma^3 \frac{\tau}{1 \frac{18G}{1 - 72L^22\gamma^2\tau} + 2\mathbb{E} \left[ \left\| v_t^{(\psi)} \right\|^2 \right] \right) \\
+ \frac{L^2\gamma^3}{2} + \frac{L^2\gamma^2}{2} \sum_{u \in K(t)} \left( \frac{4L^2\gamma^4}{n} \sum_{u' = 1}^{\psi(u)-1} \sum_{u' = 1}^{\psi(u)-1} \left( 1 - \frac{1}{n} \right)^{\phi(u) - k' - 1} \mathbb{E} \left[ \left\| \bar{v}_u^{(\psi)} \right\|^2 \right]
where (a) follows from Eq. 80, (b) uses Lemma 8, (c) follows from Eq. 77, (d) follows from Lemma 8, (e) follows from Assumption 4. Thus, we have

\[
e^{(u_{t+1} + K(t))} 
\leq (1 - \frac{\gamma \mu}{4}) e(w_t) + 2L^2 \eta_1 (L^2 \gamma^3 \tau + L \gamma^2) \left(\left(1 - \frac{1}{n}\right) \sigma(w_0) + 2\sigma(w_t)\right)
\]

\[
+ \left(2L^2 \gamma^3 \tau \frac{1}{(L^2 \gamma^2 \tau + L \gamma^2)\eta_1 L^2 \gamma^2 \tau + 8L^2 \gamma^2 \eta_1} \right)^{18G} \frac{1}{1 - 72L^2 \gamma^2 \tau}
\]

\[
+ 4(L^2 \gamma^3 \tau + L \gamma^2) \frac{L^2 \eta_1^2}{n} \sum_{c_0}^{v(u)} \left(1 - \frac{1}{n}\right)^{v(u) - k'} \sigma(w_{\bar{u}_{k'}})
\]

\[
= (1 - \frac{\gamma \mu}{4}) e(w_t) + \left(-\frac{\gamma \mu}{4} + 2c_1 + c_2\right) \sigma(w_t) + c_1 \left(1 - \frac{1}{n}\right)^{v(t)} \sigma(w_0)
\]

\[
+ c_2 \sum_{k'=1}^{v(u)} \left(1 - \frac{1}{n}\right)^{v(u) - k'} \sigma(w_{\bar{u}_{k'}}) + c_0
\]  

(83)

where \{\overline{u}_0, \overline{u}_1, \ldots, \overline{u}_{v(u) - 1}\} are the all start time counters for the global time counters from 0 to u.

We define the Lyapunov function as \(L_t = \sum_{k=0}^{v(t)} \rho^{v(t) - k} e(w_{\bar{u}_k})\) where \(\rho \in (1 - \frac{1}{n}, 1)\), we have that

\[
L_{t+1} = \rho^{v(t) + 1} e(w_0) + \sum_{k=0}^{v(t)} \rho^{v(t) - k} e(w_{\bar{u}_{k+1}})
\]

\[
\leq \rho^{v(t) + 1} e(w_0) + \sum_{k=0}^{v(t)} \rho^{v(t) - k}\left[\left(1 - \frac{\gamma \mu}{4}\right) e(w_{\bar{u}_k}) + \left(-\frac{\gamma \mu^2}{4} + 2c_1 + c_2\right) \sigma(w_{\bar{u}_k})\right]
\]

\[
+ c_1 \left(1 - \frac{1}{n}\right)^{v(t)} \sigma(w_0) + c_2 \sum_{k'=1}^{v(u)} \left(1 - \frac{1}{n}\right)^{v(u) - k'} \sigma(w_{\bar{u}_{k'}}) + c_0
\]

\[
= \rho^{v(t) + 1} e(w_0) + (1 - \frac{\gamma \mu}{4}) L_t + \sum_{k=0}^{v(t)} \rho^{v(t) - k}\left[\left(-\frac{\gamma \mu^2}{4} + 2c_1 + c_2\right) \sigma(w_{\bar{u}_k})\right]
\]

\[
+ c_1 \left(1 - \frac{1}{n}\right)^{v(t)} \sigma(w_0) + c_2 \sum_{k'=1}^{v(u)} \left(1 - \frac{1}{n}\right)^{v(u) - k'} \sigma(w_{\bar{u}_{k'}}) + \sum_{k=0}^{v(t)} \rho^{v(t) - k} c_0
\]

\[
\leq \rho^{v(t) + 1} e(w_0) + (1 - \frac{\gamma \mu}{4}) L_t + \left(-\frac{\gamma \mu^2}{4} + 2c_1 + c_2\right) \sigma(w_{\bar{u}_k}) + \frac{c_0}{1 - \rho}
\]

\[
\leq \rho^{v(t) + 1} e(w_0) + (1 - \frac{\gamma \mu}{4}) L_t - \left(\frac{\gamma \mu^2}{4} - 2c_1 - c_2\right) \frac{2L e(w_{\bar{u}_})}{1 - \rho} + \frac{c_0}{1 - \rho}
\]  

(85)

where (a) follows from Eq. 83, (b) holds by approximately choosing \(\gamma\) such that the terms related to \(\sigma(w_{\bar{u}_k})\) \((k = 0, \ldots, v(t) - 1)\) are negative, because the signs related to the lowest orders of \(\sigma(w_{\bar{u}_k})\) \((k = 0, \ldots, v(t) - 1)\) are negative. In the following we give the detailed analysis of choosing a suitable \(\gamma\) such that terms related to \(\sigma(w_{\bar{u}_k})\) \((k = 0, \ldots, v(t) - 1)\) are negative. We first consider \(k = 0\). Assume that \(C(\sigma(w_0))\) is the coefficient term of \(\sigma(w_0)\) in follows from Eq. 84, we have that

\[
C(\sigma(w_0))
\]

\[
= \rho^{v(t)} \left(-\frac{\gamma \mu^2}{4} + 2c_1 + c_2\right) + c_1 \sum_{k=0}^{v(t)} \rho^{v(t) - k} \left(1 - \frac{1}{n}\right)^k
\]
\[ \begin{align*}
&= \rho^{v(t)} \left( -\frac{\gamma \mu^2}{4} + 2c_1 + c_2 + c_1 \sum_{k=0}^{v(t)-1} \left( 1 - \frac{1}{n} \right)^k \right) \\
&\leq \rho^{v(t)} \left( -\frac{\gamma \mu^2}{4} + 2c_1 + c_2 + c_1 \frac{1}{1 - \frac{1}{2}} \right) \\
&= \rho^{v(t)} \left( -\frac{\gamma \mu^2}{4} + 2c_1 + c_2 + \left( 2 + \frac{1}{1 - \frac{1}{2 \rho}} \right) \right) \\
&\leq \rho^{v(t)} \left( -\frac{\gamma \mu^2}{4} + 2c_1 + c_2 \sum_{i=k+1}^{v(t)-1} \rho^{i-k} - \frac{1}{n} \right) \\
&= \rho^{v(t)-k} \left( -\frac{\gamma \mu^2}{4} + 2c_1 + c_2 \sum_{i=k+1}^{v(t)-1} \rho^{i-k} - \frac{1}{n} \right) \\
&\leq \rho^{v(t)-k} \left( -\frac{\gamma \mu^2}{4} + 2c_1 + c_2 \sum_{i=k+1}^{v(t)-1} \frac{1}{1 - \frac{1}{2\rho}} \right) \\
&= \rho^{v(t)-k} \left( -\frac{\gamma \mu^2}{4} + 2c_1 + c_2 \left( 1 + \frac{1}{1 - \frac{1}{2\rho}} \right) \right) \\
\end{align*} \]

Based on Eq. 86, we can carefully choose \( \gamma \) such that \(-\gamma \mu^2/4 + 2c_1 + c_2 \left( 2 + \frac{1}{1 - \frac{1}{2\rho}} \right) \leq 0 \).

Assume that \( C(\sigma(w_{\bar{u}_k})) \) is the coefficient term of \( \sigma(w_{\bar{u}_k}) \) \((k = 1, \ldots, v(t) - 1)\) in the big square brackets of follows from Eq. 77, we have that

\[ C(\sigma(w_{\bar{u}_k})) \]

\[ \begin{align*}
&= \rho^{v(t)-k} \left( -\frac{\gamma \mu^2}{4} + 2c_1 + c_2 \right) + c_2 \sum_{i=k+1}^{v(t)-1} \rho^{i-k} \left( 1 - \frac{1}{n} \right) \\
&= \rho^{v(t)-k} \left( -\frac{\gamma \mu^2}{4} + 2c_1 + c_2 \right) + c_2 \sum_{i=k+1}^{v(t)-1} \rho^{i-k} \left( 1 - \frac{1}{n} \right) \\
&\leq \rho^{v(t)-k} \left( -\frac{\gamma \mu^2}{4} + 2c_1 + c_2 \right) + c_2 \sum_{i=k+1}^{v(t)-1} \rho^{i-k} \left( 1 - \frac{1}{n} \right) \\
&= \rho^{v(t)-k} \left( -\frac{\gamma \mu^2}{4} + 2c_1 + c_2 \right) \sum_{i=k+1}^{v(t)-1} \rho^{i-k} \left( 1 - \frac{1}{n} \right) \\
\end{align*} \]

Based on Eq. 87, we can carefully choose \( \gamma \) such that \(-\gamma \mu^2/4 + 2c_1 + c_2 \left( 1 + \frac{1}{1 - \frac{1}{2\rho}} \right) \leq 0 \).

Thus, based on Eq. 84, we have that

\[ \begin{align*}
&\left( \frac{\gamma \mu^2}{4} - 2c_1 - c_2 \right) \frac{2}{L} e(w_{\bar{u}_k}) \\
&\leq \left( \frac{\gamma \mu^2}{4} - 2c_1 - c_2 \right) \frac{2}{L} e(w_{\bar{u}_k}) + \mathcal{L}_{t+|K(t)|} \\
&\leq \rho^{v(t)+1} e(w_0) + \left( 1 - \frac{\gamma \mu}{4} \right) \mathcal{L}_0 + c_0 \frac{v(t)}{1 - \rho} \\
&\leq \left( 1 - \frac{\gamma \mu}{4} \right) v(t) \mathcal{L}_0 + \rho^{v(t)+1} e(w_0) \sum_{k=0}^{v(t)-1} \left( 1 - \frac{\gamma \mu}{4} \right)^k + c_0 \frac{v(t)}{1 - \rho} \sum_{k=0}^{v(t)-1} \left( 1 - \frac{\gamma \mu}{4} \right)^k \\
&\leq \left( 1 - \frac{\gamma \mu}{4} \right) v(t) \mathcal{L}_0 + \rho^{v(t)+1} e(w_0) \left( 1 - \frac{\gamma \mu}{4} \right) + c_0 \frac{4}{1 - \rho} \gamma \mu \\
&\leq \frac{2\rho - 1 + \frac{\gamma \mu}{4}}{\rho - 1 + \frac{\gamma \mu}{4}} \rho^{v(t)+1} e(w_0) + \frac{c_0}{1 - \rho} \gamma \mu \\
\end{align*} \]

where (a) follows from Eq. 84, (b) holds by using Eq. 84 recursively, (c) uses the fact that \( 1 - \frac{\gamma \mu}{4} \leq \rho \). According to Eq. 88, we have that

\[ e(w_{u_{\bar{u}_k}}) \leq \frac{2\rho - 1 + \frac{\gamma \mu}{4}}{\rho - 1 + \frac{\gamma \mu}{4}} \rho^{v(t)+1} e(w_0) + \frac{c_0}{1 - \rho} \gamma \mu (1 - \rho) \left( \frac{\gamma \mu}{4} - 2c_1 - c_2 \right) \]
Thus, under , to obtain the accuracy $\epsilon$ of Problem $\square$, for VFB$^2$-SAGA, we can carefully choose $\gamma$ such that

$$1 - 72L^2\gamma^2\tau > 0$$  \hfill (89)

$$\frac{4c_0}{\gamma\mu(1 - \rho)\left(\frac{2\mu^2}{4} - 2c_1 - c_2\right)} \leq \frac{\epsilon}{2}$$  \hfill (90)

$$0 < 1 - \frac{\gamma\mu}{4} < 1$$  \hfill (91)

$$-\frac{\gamma^2\mu^2}{4} + 2c_1 + c_2 \left(1 + \frac{1}{1 - \frac{1 - \frac{1}{\rho}}{\rho}}\right) \leq 0$$  \hfill (92)

$$-\frac{\gamma^2\mu^2}{4} + c_2 + c_1 \left(2 + \frac{1}{1 - \frac{1}{\rho}}\right) \leq 0$$  \hfill (93)

and let

$$\frac{2\rho - 1 + 2\mu}{(\rho - 1 + 2\mu)(\frac{2\mu^2}{4} - 2c_1 - c_2)} \rho^{\psi(t) + 1}w_0 \leq \frac{\epsilon}{2},$$

we have that

$$\nu(t) \geq \frac{\log \rho}{\log \frac{\epsilon}{2}}$$  \hfill (95)

This completes the proof. \hfill $\square$

## G Convergence Analyses of Nonconvex Problems

### Proof of Theorem 4

**Lemma 10.** For all $t$ (whether the $t$-th global iteration is a dominated or collaborative update), there is

$$\sum_{t=0}^{S-1} \mathbb{E}\|\tilde{v}_t^{\psi(t)}\|^2 \leq \frac{4}{1 - \lambda_1} \sum_{t=0}^{S-1} \mathbb{E}\|\hat{v}_t^{\psi(t)}\|^2,$$  \hfill (96)

where $S$ denotes the total number of iterations, $\lambda_1 = 6L^2\gamma^2\tau$.

**Proof of Lemma 10:** First, when the $t$-th global iteration corresponds to collaborative update, we have

$$\mathbb{E}\|\tilde{v}_t^{\psi(t)}\|^2 = \mathbb{E}\|\nabla \theta(x_t)\bar{g}(\bar{x}_t) + \nabla g_v(x_t)g((\bar{x}_t)\bar{g}(\bar{x}_t))\|^2$$

$$= \mathbb{E}\|\nabla \theta(x_t)\bar{g}(\bar{x}_t) + \nabla g_v(x_t)g(\bar{x}_t)g((\bar{x}_t)\bar{g}(\bar{x}_t)) + \nabla g_v(x_t)g((\bar{x}_t)\bar{g}(\bar{x}_t))\|^2$$

$$\leq (a) 2\mathbb{E}\|\tilde{v}_t^{\psi(t)}\|^2 + 2\mathbb{E}\|\nabla g_v(x_t)g(\bar{x}_t)g((\bar{x}_t)\bar{g}(\bar{x}_t)) - \nabla g_v(x_t)g((\bar{x}_t)\bar{g}(\bar{x}_t))\|^2$$

$$\leq (b) 2\mathbb{E}\|\tilde{v}_t^{\psi(t)}\|^2 + 2L_g\mathbb{E}\|\bar{x}_t\bar{g}(\bar{x}_t) - (\bar{x}_t)\bar{g}(\bar{x}_t)\|^2$$

$$\leq (c) 2\mathbb{E}\|\tilde{v}_t^{\psi(t)}\|^2 + 2L_g^2\mathbb{E}\| \bar{x}_t\bar{g}(\bar{x}_t) - (\bar{x}_t)\bar{g}(\bar{x}_t)\|^2$$

$$\leq (d) 2\mathbb{E}\|\tilde{v}_t^{\psi(t)}\|^2 + 2L_g^2\mathbb{E}\| \sum_{t' \in D^*(t), \psi(t') = \psi(t)} \bar{v}_{t'}^{\psi(t')}\|^2$$

$$\leq (e) 2\mathbb{E}\|\tilde{v}_t^{\psi(t)}\|^2 + 2L_g^2\mathbb{E}\| \sum_{t' \in D^*(t)} \bar{v}_{t'}^{\psi(t')}\|^2$$  \hfill (97)

where (a) follows from $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, (b) follows from Assumption 2, (c) follows from the Eq. 5, (d) follows from
where (a) follows from Assumption 3 and (b) follows from Assumption 2, (c) follows from the Eq. 5, (d) follows from the definition of $L_\star$. Combining Eqs. 97 and 98 we have

$$\mathbb{E}\|\bar{v}_t^{\psi(t)}\|^2 \leq 4\mathbb{E}\|\bar{v}_t^{\psi(t)}\|^2 + 6L_\star^2\gamma^2\tau_2 \sum_{t'\in D'(t)} \mathbb{E}\|\bar{v}_{t'}^{\psi(t')}\|^2$$

(99)

Summing Eq. (99) for all iterations (assume the number of total iterations is $S$), there is

$$\sum_{t=0}^{S-1} \mathbb{E}\|\bar{v}_t^{\psi(t)}\|^2 \leq 4\sum_{t=0}^{S-1} \mathbb{E}\|\bar{v}_t^{\psi(t)}\|^2 + 6L_\star^2\gamma^2\tau_2 \sum_{t=0}^{S-1} \sum_{t'\in D'(t)} \mathbb{E}\|\bar{v}_{t'}^{\psi(t')}\|^2$$

$$\leq 4\sum_{t=0}^{S-1} \mathbb{E}\|\bar{v}_t^{\psi(t)}\|^2 + 6L_\star^2\gamma^2\tau \sum_{t=0}^{S-1} \mathbb{E}\|\bar{v}_t^{\psi(t)}\|^2,$$

(100)

where (a) follows from Assumption 3. When the $t$-th global iteration corresponds to dominated update, it is obviously that

$$\sum_{t=0}^{S-1} \mathbb{E}\|\bar{v}_t^{\psi(t)}\|^2 = \sum_{t=0}^{S-1} \mathbb{E}\|\bar{v}_t^{\psi(t)}\|^2 < 4\sum_{t=0}^{S-1} \mathbb{E}\|\bar{v}_t^{\psi(t)}\|^2 + 6L_\star^2\gamma^2\tau \sum_{t=0}^{S-1} \mathbb{E}\|\bar{v}_t^{\psi(t)}\|^2$$

(101)

Combining Eqs. (100) and (101), there is

$$\sum_{t=0}^{S-1} \mathbb{E}\|\bar{v}_t^{\psi(t)}\|^2 < 4\sum_{t=0}^{S-1} \mathbb{E}\|\bar{v}_t^{\psi(t)}\|^2 + 6L_\star^2\gamma^2\tau \sum_{t=0}^{S-1} \mathbb{E}\|\bar{v}_t^{\psi(t)}\|^2$$

(102)

whether the $t$-th global iteration corresponds to collaborative update or dominated one, which implies that if $1 - \lambda_1 > 0$ there is

$$\sum_{t=0}^{S-1} \mathbb{E}\|\bar{v}_t^{\psi(t)}\|^2 < \frac{4}{1 - 6L_\star^2\gamma^2\tau} \sum_{t=0}^{S-1} \mathbb{E}\|\bar{v}_t^{\psi(t)}\|^2 = \frac{4}{1 - \lambda_1} \sum_{t=0}^{S-1} \mathbb{E}\|\bar{v}_t^{\psi(t)}\|^2$$

(103)

where $\lambda_1 = 6L_\star^2\gamma^2\tau$, this completes the proof. 

\[\Box\]

**Lemma 11.** For $\forall t$ (whether the $t$-th global iteration is a dominated or collaborative update), there is

$$\mathbb{E}\|v_t^{\psi(t)} - \bar{v}_t^{\psi(t)}\|^2 \leq 2L_\star^2\gamma^2\tau_1 \sum_{t'\in D(t)} \mathbb{E}\|\bar{v}_{t'}^{\psi(t')}\|^2 + 8L_\star^2\gamma^2\tau_2 \sum_{t'\in D'(t)} \mathbb{E}\|\bar{v}_{t'}^{\psi(t')}\|^2.$$
Proof of Lemma \cite{17}. First, we give the bound of $\mathbb{E}\|\tilde{v}_t^{(t)} - \tilde{v}_t^{(t)}\|^2$ as follow

\begin{align}
\mathbb{E}\|\tilde{v}_t^{(t)} - \tilde{v}_t^{(t)}\|^2 &\leq (a) \mathbb{E}\|\nabla_{\tilde{g}(t)} f(\tilde{w}_t) - \nabla_{\tilde{g}(t)} f(\tilde{w}_t) + \nabla_{\tilde{g}(t)} g(\tilde{w}_t)\|_2^2
+ \mathbb{E}\|\nabla_{\tilde{g}(t)} g(\tilde{w}_t)\|_2^2
+ \mathbb{E}\|\nabla_{\tilde{g}(t)} g(\tilde{w}_t)\|_2^2

&\leq (c) 2\mathbb{E}\|\nabla_{\tilde{g}(t)} f_t(\tilde{w}_t) - \nabla_{\tilde{g}(t)} f_t(\tilde{w}_t)\|^2 + 2\mathbb{E}\|\nabla_{\tilde{g}(t)} g(\tilde{w}_t)\|_2^2
+ (d) 2\mathbb{E}\|\tilde{w}_t - \tilde{v}_t\|^2 + 2\mathbb{E}\|\tilde{v}_t - (\tilde{w}_t)\|_2^2

&\leq (e) 2L^2\gamma^2\mathbb{E}\|\tilde{v}_t - \tilde{v}_t\|^2 + 2L^2\gamma^2\mathbb{E}\|\tilde{v}_t - \tilde{v}_t\|^2
+ (f) 2L^2\gamma^2\mathbb{E}\||\tilde{v}_t - \tilde{v}_t\||^2 + 2L^2\gamma^2\mathbb{E}\||\tilde{v}_t - \tilde{v}_t\||^2

&\leq (g) 4L^2\gamma^2\mathbb{E}\|\tilde{v}_t - \tilde{v}_t\|^2
\end{align}

(105)

where (a) follows from the definition of $\tilde{v}_t^{(t)}$ and the definitions of $\tilde{v}_t^{(t)}$ in different type of updates (dominated or collaborative one), (b) follows from $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, (c) follows from the definition of $\tilde{v}_t^{(t)}$ and $\tilde{v}_t^{(t)}$, (d) follows from Assumptions \cite{2}, (e) follows from Eqs. \cite{4} and \cite{5}, inequalities, (f) follows from Assumptions \cite{3} to \cite{3} and $\sum_{i=1}^n a_i \| \leq n \sum_{i=1}^n \| a_i \|^2$, (g) follows from the definition of $L_s$. Then we consider the bound

\begin{align}
\mathbb{E}\|v_t^{(t)} - \bar{v}_t^{(t)}\|^2 &\leq (a) \mathbb{E}\|v_t^{(t)} - \bar{v}_t^{(t)} + \bar{v}_t^{(t)} - \bar{v}_t^{(t)}\|^2

&\leq (b) \mathbb{E}\|\nabla_{\tilde{g}(t)} f_t(w_t) - \nabla_{\tilde{g}(t)} f_t(\tilde{w}_t)\|^2 + 2\mathbb{E}\|\bar{v}_t^{(t)} - \bar{v}_t^{(t)}\|^2

&\leq (c) 2L^2\gamma^2\mathbb{E}\|\tilde{v}_t - \tilde{v}_t\|^2 + 2\mathbb{E}\|\tilde{v}_t - \tilde{v}_t\|^2

&\leq (d) 2L^2\gamma^2\mathbb{E}\|\tilde{v}_t - \tilde{v}_t\|^2 + 2\mathbb{E}\|\tilde{v}_t - \tilde{v}_t\|^2

&\leq (e) 2L^2\gamma^2\mathbb{E}\|\tilde{v}_t - \tilde{v}_t\|^2
\end{align}

(106)

where (a) follows from $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, (b) follows from Assumptions \cite{2}, (c) follows from Eqs. \cite{4} and \cite{5} inequalities, (d) follows from Assumptions \cite{3} to \cite{3} and $\sum_{i=1}^n a_i \| \leq n \sum_{i=1}^n \| a_i \|^2$, (e) follows from the definition of $L_s$ and Eq. \cite{105}. This completes the proof.

Lemma 12. Assuming $S = qc$, where $c > 0$ is an integer, we have

\begin{align}
\sum_{i = 1}^{S-1} \mathbb{E}\|\nabla f(w_i)\|^2 &\leq (a) 2L^2\gamma^2\mathbb{E}\|\tilde{v}_t^{(t)}\|^2 + 2\sum_{u=0}^{S-1} \mathbb{E}\|\nabla_{\tilde{g}(u)} f(w_u)\|^2
\end{align}

(107)

Proof of Lemma \cite{17}. Given $t$ denotes a global iteration number, if the $t$-th global iteration is a dominated one, then for any $t' \in K'(t)$, there is

\begin{align}
\mathbb{E}\|\nabla_{\tilde{g}(t')} f(w_t)\|^2 &\leq (a) \mathbb{E}\|\nabla_{\tilde{g}(t')} f(w_t) - \nabla_{\tilde{g}(t')} f(w_t) + \nabla_{\tilde{g}(t')} f(w_t)\|^2

&\leq (a) 2\mathbb{E}\|\nabla_{\tilde{g}(t')} f(w_t) - \nabla_{\tilde{g}(t')} f(w_t)\|^2 + 2\mathbb{E}\|\nabla_{\tilde{g}(t')} f(w_t)\|^2

&\leq (b) 2L^2\gamma^2\mathbb{E}\|\tilde{w}_t - w_{t'}\|^2 + 2\mathbb{E}\|\nabla_{\tilde{g}(t')} f(w_t)\|^2
\end{align}

\[ \]
\[
\begin{aligned}
\text{(c)} & \leq 2L^2\gamma^2E\sum_{u \in \{t, \ldots, t'\}} U_{\psi(u)}\tilde{v}_{u}^\psi(u))^2 + 2E\|\nabla\varphi_{\psi(t')}f(w_{t'})\|^2 \\
\text{(d)} & \leq 2L^2\gamma^2\eta_2 \sum_{u \in \{t, \ldots, t'\}} E\|\tilde{v}_{u}^\psi(u))^2 + 2E\|\nabla\varphi_{\psi(t')}f(w_{t'})\|^2 
\end{aligned}
\] (108)

where (a) follows from \(\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2\), (b) follows from Assumptions 2, (c) follows from Eq. 4, and (d) follows from the bound of \(|K'(t)|\) and \(\|\sum_{i=1}^{n} a_i\|^2 \leq n \sum_{i=1}^{n} \|a_i\|^2\). Summing above for \(t' \in K'\) and all \(t \in A(S)\) we have

\[
\begin{aligned}
\sum_{t \in A(S)} \sum_{t' \in K'(t)} E\|\nabla\varphi_{\psi(t')}f(w_{t})\|^2 &\leq 2L^2\gamma^2\eta_2 \sum_{t \in A(S)} \sum_{t' \in K'(t)} E\|\tilde{v}_{t}^\psi(t))^2 + 2 \sum_{t \in A(S)} E\|\nabla\varphi_{\psi(t')}f(w_{t'})\|^2 \\
&\leq 2L^2\gamma^2\eta_2 \sum_{t=0}^{S-1} E\|\tilde{v}_{t}^\psi(t))^2 + 2 \sum_{t=0}^{S-1} E\|\nabla\varphi_{\psi(t')}f(w_{t'})\|^2
\end{aligned}
\] (109)

where (a) follows from the bound of \(|K'(t)|\) and the definition of A(S). For \(\forall u \in A(S)\) there is

\[
E\|\nabla f(w_u)\|^2 = \|\sum_{u' \in K'(u)} \nabla\varphi_{\psi(u')}f(w_u)\|^2 \leq \sum_{u' \in K'(u)} E\|\tilde{v}_{t}^\psi(t))^2 + 2 \sum_{t \in A(S)} E\|\nabla\varphi_{\psi(t')}f(w_{t'})\|^2
\] (110)

where (a) follows from the orthogonality between all coordinates. Combing Eqs. 109 and 110 there is

\[
\sum_{u \in A(S)} E\|\nabla f(w_u)\|^2 \leq 2L^2\gamma^2\eta_2 \sum_{t=0}^{S-1} E\|\tilde{v}_{t}^\psi(t))^2 + 2 \sum_{t=0}^{S-1} E\|\nabla\varphi_{\psi(t')}f(w_{t'})\|^2
\] (111)

This completes the proof.

\[\square\]

**Proof of Theorem 7.** For \(\forall t\) denotes a global iteration, we have that

\[
\begin{aligned}
E f(w_{t+1}) &\leq E f(w_{t}) + \langle \nabla f(w_{t}), w_{t+1} - w_{t} \rangle + \frac{L}{2} \|w_{t+1} - w_{t}\|^2 \\
&\leq \frac{L}{2} \|w_{t+1} - w_{t}\|^2 + \|v_{t} - \tilde{v}_{t}\|^2 + \gamma \|\nabla f(w_{t})\|^2 \\
&\leq \frac{L}{2} \|w_{t+1} - w_{t}\|^2 + \|v_{t} - \tilde{v}_{t}\|^2 + \gamma \|\nabla f(w_{t})\|^2 \\
&\leq \frac{L}{2} \|w_{t+1} - w_{t}\|^2 + \|v_{t} - \tilde{v}_{t}\|^2 + \gamma \|\nabla f(w_{t})\|^2 \\
\end{aligned}
\] (112)

where the inequalities (a) follows form Assumption 1, (b) follows from that \(\tilde{v}_{t} = \nabla\varphi_{\psi(t)} f_{t}(w_{t})\) for a specific party, (c) follows from \(\langle a, b \rangle \leq \frac{1}{2} \|a\|^2 + \|b\|^2\), (d) follows from Lemma 11 and the definition of \(L_{\gamma}\). Summing Eq. 112 over all
0 \leq t \leq S - 1, we obtain
\[
\frac{\gamma}{2} \sum_{t=0}^{S-1} \mathbb{E} \left\| \nabla \varphi(w_t) \right\|^2 \\
\leq \sum_{t=0}^{S-1} \mathbb{E} \left[ f(w_t) - f(w_{t+1}) \right] + L^2 \gamma^2 \sum_{t=0}^{S-1} \mathbb{E} \left\| \nabla \psi(t) \right\|^2 \\
+ 4L^2 \gamma^2 \sum_{t=0}^{S-1} \mathbb{E} \left\| \nabla \psi(t) \right\|^2 + \frac{L^2 \gamma}{2} \sum_{t=0}^{S-1} \mathbb{E} \left\| \nabla \psi(t) \right\|^2 
\]
Note that \( \nabla \varphi(w_t) \) is the gradient of coordinate \( \psi(t) \), while to obtain the global convergence rate it is necessary to focus on the gradient of all coordinates \( \nabla f(w_t) \). Combining Eq. 113 with Lemma 12, there is
\[
\sum_{u \in \mathcal{A}(S)} \mathbb{E} \left\| \nabla f(w_u) \right\|^2 \leq \left( \frac{4}{S} \right) \sum_{t=0}^{S-1} \mathbb{E} \left[ f(w^0) - f(w^*) \right] + \left( 2L^2 \gamma^2 \tau + 2L \gamma \right) \sum_{t=0}^{S-1} \mathbb{E} \left\| \nabla \psi(t) \right\|^2 \\
\leq \left( \frac{4}{S} \right) \sum_{t=0}^{S-1} \mathbb{E} \left[ f(w^0) - f(w^*) \right] + \left( 2L^2 \gamma^2 \tau + 2L \gamma \right) \sum_{t=0}^{S-1} \mathbb{E} \left\| \nabla \psi(t) \right\|^2 \\
\leq \left( \frac{4}{S} \right) \sum_{t=0}^{S-1} \mathbb{E} \left[ f(w^0) - f(w^*) \right] + \left( 2L^2 \gamma^2 \tau + 2L \gamma \right) \frac{4}{1 - 6L^2 \gamma^2 \tau} G
\]
where (a) follows from Assumptions 3 and 3, (b) follows from the definition of \( \tau \), (c) follows from Lemma 10 and 10. Which implies that
\[
\frac{1}{S} \sum_{u \in \mathcal{A}(S)} \mathbb{E} \left\| \nabla f(w_u) \right\|^2 \leq \frac{4}{S} \sum_{t=0}^{S-1} \mathbb{E} \left[ f(w^0) - f(w^*) \right] + \left( 2L^2 \gamma^2 \tau + 2L \gamma \right) \frac{4}{1 - 6L^2 \gamma^2 \tau} G
\]
Note that for \( S = q \) and \( c \) is an integer, there is \( |A(S)| = \frac{S}{q} \), and then we have
\[
\frac{1}{|A(S)|} \sum_{u \in \mathcal{A}(S)} \mathbb{E} \left\| \nabla f(w_u) \right\|^2 \leq \frac{4}{|A(S)|} \sum_{t=0}^{S-1} \mathbb{E} \left[ f(w^0) - f(w^*) \right] + \left( 2L^2 \gamma^2 \tau + 2L \gamma \right) \frac{4}{1 - 6L^2 \gamma^2 \tau} G_1
\]
where \( G_1 = qG \). To obtain the \( \epsilon \)-first-order stationary solution one can choose suitable \( \gamma \), such that
\[
1 - 6L^2 \gamma^2 \tau > 0
\]
\[
\frac{4}{|A(S)|} \sum_{t=0}^{S-1} \mathbb{E} \left[ f(w^0) - f(w^*) \right] + \left( 2L^2 \gamma^2 \tau + 2L \gamma \right) \frac{4}{1 - 6L^2 \gamma^2 \tau} G_1 \leq \frac{\epsilon}{2}
\]
which implies that if \( \tau \) is upper bounded, i.e., \( \tau \leq \frac{512G}{3L^2} \) (one can obtain this by combining Eqs. 117 and 119 and assuming Eq. 118 holds), we can carefully choose the stepsize as
\[
\gamma = \frac{\epsilon}{32L \gamma G}
\]
and if the total epochs number (i.e., \( v'(S) \)) of \( S \) global iterations denoted as \( T \) satisfying
\[
T \geq \frac{256L \gamma qGE(f(w_0) - f(w^*))}{\epsilon^2}
\]
the \( \epsilon \)-first-order stationary solution is obtained:
\[
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\| \nabla f(w_t) \right\|^2 \leq \epsilon
\]
this completes the proof.
\[
\square
\]
Proof of Theorem 5

Lemma 13. For all outer loop $s = 1, \cdots, S$ we define $A'(s)$ as all epoches during this outer loop, there is

$$\sum_{u \in A'(s)} \sum_{t \in K'} \mathbb{E}[\|v^t_u\|_2^2] \leq \lambda_\gamma \sum_{u \in A'(s)} \sum_{t \in K'} \mathbb{E}[\|v^t_u\|_2^2],$$

where $\lambda_\gamma = \frac{2}{1 - 20L^2\gamma^2\tau} > 0$.

Proof of Lemma 13. First, we have

$$\mathbb{E}[\|v^t_u\|_2^2] = \mathbb{E}[\|v^t_u - v^t_0 + v^t_0\|_2^2] \leq 2\mathbb{E}[\|v^t_u - v^t_0\|_2^2] + 2\mathbb{E}[\|v^t_0\|_2^2]$$

where (a) follows from $\|a + b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$. (b) follows from Lemma 3. Summing Eq. (123) over an outer loop $s$, then there is

$$\sum_{u \in A'(s)} \sum_{t \in K'} \mathbb{E}[\|v^t_u\|_2^2] \leq \lambda_\gamma \sum_{u \in A'(s)} \sum_{t \in K'} \mathbb{E}[\|v^t_u\|_2^2]$$

This completes the proof.

Proof of Theorem 5. Similar to the proof of Theorem 4, we first apply Lemma 3 to an epoch (or an outer loop) $s$, and there is

$$\sum_{u \in A'(s)} \mathbb{E}[\|\nabla f(w^s_u)\|_2^2] \leq \sum_{u \in A'(s)} \sum_{t \in K'} \mathbb{E}[\|v^t_u\|_2^2] + 2\sum_{u \in A'(s)} \sum_{t \in K'} \mathbb{E}[\|\nabla g_{\varphi(t)}f(w^t_u)\|_2^2]$$

Summing Eq. (126) over outer loops $1, \cdots, S$, we have

$$\sum_{s=1}^S \sum_{u \in A'(s)} \sum_{t \in K'} \mathbb{E}[\|\nabla g_{\varphi(t)}f(w^s_u)\|_2^2] \leq 2\sum_{s=1}^S \sum_{u \in A'(s)} \sum_{t \in K'} \mathbb{E}[\|v^t_u\|_2^2]
+ 2\sum_{s=1}^S \sum_{u \in A'(s)} \sum_{t \in K'} \mathbb{E}[\|\nabla g_{\varphi(t)}f(w^s_u)\|_2^2]$$

Then we bound R.H.S. as follow. First, we consider the bound of $\mathbb{E}[\|v^t_u\|_2^2]$, and define

$$\zeta^s_t = \nabla g_{\varphi(t)}f(w^s_t) - \nabla g_{\varphi(t)}f(w^s_t)$$

where $w^s_t$ denotes $w^s_t$ at outer loop $s$. From the definition of $v^t_u$, one can get:

$$\mathbb{E}[\|v^t_u\|_2^2] = \mathbb{E}[\|\zeta^s_t + \nabla g_{\varphi(t)}f(w^s_t)\|_2^2]
\leq 2\mathbb{E}[\|\nabla g_{\varphi(t)}f(w^s_t)\|_2^2] + 2\mathbb{E}[\|\zeta^s_t\|_2^2]$$

where (a) follows from $\|a + b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$, and $\mathbb{E}[\zeta^s_t] = \nabla f(w^s_t) - \nabla f(w^s_t)$. From the above equality, we have

$$\mathbb{E}[\|v^t_u\|_2^2] \leq 2\mathbb{E}[\|\nabla g_{\varphi(t)}f(w^s_t)\|_2^2] + 2\mathbb{E}[\|\nabla g_{\varphi(t)}f(w^s_t) - \nabla g_{\varphi(t)}f(w^s_t)\|_2^2]$$

(129)
where (a) follows from that $\mathbb{E}[\zeta - E[\zeta]]^2 \leq \mathbb{E}[\zeta]^2$, (b) follows from Assumption 2. We define $\tilde{\nabla}^*_{\varphi(t)} = \nabla_{\varphi(t)} L(\bar{u}^*_t) + \nabla_{\varphi(t)} g((\bar{w}^*_t))g_{\varphi(t)} = \theta_1 \cdot (x_t)g_{\varphi(t)} + \nabla_{\varphi(t)} g((\bar{w}^*_t))g_{\varphi(t)}$ when the $t$-th global iteration denotes a collaborative update, while $\tilde{\nabla}^*_{\varphi(t)} = \nabla_{\varphi(t)} f(\bar{w}_t)$ if a dominated update. Then we derive the upper bound of $\mathbb{E}[w^*_{t+1} - w^*]^2$

\[
\mathbb{E}[w^*_{t+1} - w^*]^2 = \mathbb{E}[w^*_{t+1} - w^* + w^* - w^*]^2 - 2\mathbb{E}(w^*_{t+1} - w^*, w^* - w^*) \\
= \gamma^2 \mathbb{E}[\tilde{\nabla}^*(t)]^2 + \mathbb{E}[w^* - w^*]^2 - 2\gamma \mathbb{E}\left(\tilde{\nabla}^*_{\varphi(t)}, w^* - w^*\right) \\
\leq \gamma^2 \mathbb{E}[||\tilde{\nabla}^*(t)||^2 + ||w^* - w^*||^2 + 2\gamma \mathbb{E}\left(\frac{1}{2\beta_t}||\tilde{\nabla}^*_{\varphi(t)}||^2 + \frac{\beta_t}{2}||w^* - w^*||^2\right) \\
= \gamma^2 \mathbb{E}[||\tilde{\nabla}^*(t)||^2 + \frac{\gamma}{\beta_t} \mathbb{E}[||\tilde{\nabla}^*_{\varphi(t)}||^2 + (1 + \gamma\beta_t)\mathbb{E}[||w^* - w^*||^2]
\]

where (a) follows from Yong-Equation. For $\forall t \in K'(u)$, where $u \in A'(s)$ there is

\[
\mathbb{E}[f(w^*_{t+1}) \leq \mathbb{E}\left[f(w^*_t) + \langle \nabla f(w^*_t), w^*_{t+1} - w^*_t \rangle + \frac{L}{2}||w^*_{t+1} - w^*_t||^2\right] \\
= \mathbb{E}[f(w^*_t) - \gamma \mathbb{E}\left[\nabla f(w^*_t), \tilde{\nabla}^*_{\varphi(t)}\right] + \frac{\gamma^2 L}{2} \mathbb{E}[||\tilde{\nabla}^*(t)||^2] \\
\leq \mathbb{E}[f(w^*_t) - \frac{\gamma}{2} \mathbb{E}[||\nabla_{\varphi(t)} f(w^*_t)||^2 + ||\tilde{\nabla}^*_{\varphi(t)}||^2] \\
- \frac{||\nabla_{\varphi(t)} f(w^*_t) - \tilde{\nabla}^*_{\varphi(t)}||^2}{2} + \frac{\gamma^2 L}{2} \mathbb{E}[||\tilde{\nabla}^*(t)||^2]
\]

where (a) follows from Assumption 2 and (b) follows form $a, b = ||a||^2 + ||b||^2 - ||a - b||^2$. Next, we give the upper bound of the term $\mathbb{E}[||\nabla_{\varphi(t)} f(w^*_t) - \tilde{\nabla}^*_{\varphi(t)}||^2$:

\[
\mathbb{E}[||\nabla_{\varphi(t)} f(w^*_t) - \tilde{\nabla}^*_{\varphi(t)}||^2 \leq 2L^2\gamma^2 \tau_1 \sum_{u \in U'(t)} \mathbb{E}[||v_{\varphi(u)}^*(u)||^2 + 8L^2\gamma^2 \tau_2 \sum_{u \in U'(t)} \mathbb{E}[||v_{\varphi(u)}^*(u)||^2]
\]

Above result can be obtained by applying Lemma 3 with $v_{\varphi(t)}^*$ and $\tilde{v}_{\varphi(t)}^*$ defined in SVRG-based algorithm. From (132) and (133), it is easy to derive the following inequality:

\[
\mathbb{E}[f(w^*_{t+1}) \leq \mathbb{E}[f(w^*_t) - \frac{\gamma}{2} \mathbb{E}[||\nabla_{\varphi(t)} f(w^*_t)||^2 - \frac{\gamma}{2} \mathbb{E}[||\tilde{\nabla}^*_{\varphi(t)}||^2 + \frac{\gamma^2 L}{2} \mathbb{E}[||\tilde{\nabla}^*(t)||^2
\]

\[
+ L^2\gamma^3 \left(\tau_1 \sum_{u \in U'(t)} \mathbb{E}[||v_{\varphi(u)}^*(u)||^2 + 4\tau_2 \sum_{u \in U'(t)} \mathbb{E}[||v_{\varphi(u)}^*(u)||^2\right)
\]

Similar to many convergence analyses of nonconvex optimization, we define the Lyapunov function as

\[
R^*_t = \mathbb{E}\left[f(w^*_t) + c_t ||w^*_t - w^*||^2\right],
\]

then there is

\[
R^*_{t+1} = \mathbb{E}\left[f(w^*_{t+1}) + c_{t+1} ||w^*_{t+1} - w^*||^2\right]
\]

\[
\leq \mathbb{E}[f(w^*_t) - \frac{\gamma}{2} \mathbb{E}[||\nabla_{\varphi(t)} f(w^*_t)||^2 - \frac{\gamma}{2} \mathbb{E}[||\tilde{\nabla}^*_{\varphi(t)}||^2 + \frac{\gamma^2 L}{2} \mathbb{E}[||\tilde{\nabla}^*(t)||^2
\]

\[
+ L^2\gamma^3 \left(\tau_1 \sum_{u \in U'(t)} \mathbb{E}[||v_{\varphi(u)}^*(u)||^2 + 4\tau_2 \sum_{u \in U'(t)} \mathbb{E}[||v_{\varphi(u)}^*(u)||^2\right) \\
+ c_{t+1} \left[\gamma^2 \mathbb{E}[||v_{\varphi(t)}^*(u)||^2 + \frac{\gamma}{\beta_t} \mathbb{E}[||\tilde{\nabla}^*_{\varphi(t)}||^2 + (1 + \gamma\beta_t)\mathbb{E}[||w^*_t - w^*||^2] \right]
\]

\[
\leq \mathbb{E}[f(w^*_t) - \frac{\gamma}{2} \mathbb{E}[||\nabla_{\varphi(t)} f(w^*_t)||^2 - \frac{\gamma}{2} - \frac{c_{t+1} \gamma}{\beta_t} \mathbb{E}[||\tilde{\nabla}^*_{\varphi(t)}||^2
\]
+ L_s^2 \gamma^3 \tau_1 \sum_{t' \in D(t)} \mathbb{E}\|\tilde{v}^{\psi(t')}\|^2 + \left(\frac{\gamma^2 L}{2} + c_{t+1} \gamma^2\right)\mathbb{E}\|\tilde{u}_t^{\psi(t)}\|^2 \\
+ 4L_s^2 \gamma^3 \tau_2 \sum_{t' \in D(t)} \mathbb{E}\|\tilde{v}^{\psi(t')}\|^2 + c_{t+1} (1 + \gamma \beta_t) \mathbb{E}\|u_t^n - w^s\|^2
\end{align*}

where (a) follow from Assumptions 3 to 3 and assuming \( \frac{1}{2} \geq \frac{c_{u_t+1}}{\beta_{u_t}} \), (b) follows from Lemma 3. Denote \( 10L_s^2 \gamma^3 \tau + \gamma^2 L_s + 2c_{u_t+1} \gamma^2 \) as \( \lambda_{u_t} \), we have

\begin{align*}
\sum_{s=1}^{S} u \in \mathcal{A}(s) \sum_{t \in K'(u)} R_{u_t+1}^s \quad (137)
\end{align*}

Summing this over an outer loop \( s \) one can obtain:

\begin{align*}
\sum_{s=1}^{S} u \in \mathcal{A}(s) \sum_{t \in K'(u)} R_{u_t+1}^s \quad (138)
\end{align*}

where (a) follows from Eqs. (131) and (134). Summing this over an outer loop \( s \) one can obtain:

\begin{align*}
\sum_{s=1}^{S} u \in \mathcal{A}(s) \sum_{t \in K'(u)} R_{u_t+1}^s \quad (139)
\end{align*}

where (a) follows from Assumptions 3 to 3 and assuming \( \frac{1}{2} \geq \frac{c_{u_t+1}}{\beta_{u_t}} \), (b) follows from Lemma 3. Denote \( 10L_s^2 \gamma^3 \tau + \gamma^2 L_s + 2c_{u_t+1} \gamma^2 \) as \( \lambda_{u_t} \), we have

\begin{align*}
\sum_{s=1}^{S} u \in \mathcal{A}(s) \sum_{t \in K'(u)} R_{u_t+1}^s \quad (140)
\end{align*}

where (a) follows from Eqs. (131) and (134). Summing this over an outer loop \( s \) one can obtain:
where (a) follows from the definition of $L_s$ and $L_*$. Then we return to Eq. [127]

\[ \sum_{s=1}^{S} \sum_{u \in A'(s) \in K'(u)} \mathbb{E}\|\nabla \mathcal{G}_{\psi(u_s)} f(w_{u_s}^s)\|^2 \]

where (a) follows from Eq. [130] and $u_0$ denotes the start iteration during epoch $u$. This implies that

\[ \frac{\gamma}{2} - \lambda \gamma \lambda_{u_t} \sum_{s=1}^{S} \sum_{u \in A'(s) \in K'(u)} \mathbb{E}\|\nabla \mathcal{G}_{\psi(u_s)} f(w_{u_s}^s)\|^2 \]

where (a) follows from the definition of $L_*$. Combining Eq. [143] with [140] we have

\[ \frac{\gamma}{2} - \lambda \gamma \lambda_{u_t} \sum_{s=1}^{S} \sum_{u \in A'(s) \in K'(u)} \mathbb{E}\|\nabla \mathcal{G}_{\psi(u_s)} f(w_{u_s}^s)\|^2 \]

Rearrange Eq. [144] we have

\[ \sum_{s=1}^{S} \sum_{u \in A'(s) \in K'(u)} \mathbb{E}\|\nabla f(w_{u_t})\|^2 - \frac{\gamma}{2} - \lambda \gamma \lambda_{u_t} \sum_{s=1}^{S} \sum_{u \in A'(s) \in K'(u)} \mathbb{E}\|\nabla \mathcal{G}_{\psi(u_s)} f(w_{u_s}^s)\|^2 \]
\[ c_{u_t} = c_{u_{t+1}} (1 + \gamma \beta_{u_t}) + \frac{\gamma}{2} L_{u_t}^2 \]  
\[ \Gamma_{u_t} = \frac{\frac{\gamma}{2} - \frac{2}{1-20L^2_{u_t}\gamma^2} (10L^2_{u_t}\gamma^3 - \gamma^2 L_{u_t} + 2c_{u_{t+1}}\gamma^2)}{2 + 4\lambda_{u_t} L_{u_t}^2 L_{u_t}^2} \]  
Denote the subscript of the last iteration in \( s \)-th outer loop as \( c_s \) and set it as 0, and set  
\[ w^{s+1} = w^{s}_{u_t} = \bar{s} \]  
then there is  
\[ R^{s}_{u_t} = \mathbb{E} f(w^{s}_{u_t}) = \mathbb{E} f(w^{s+1}) \]  
Applying these to [40] we can get,  
\[ \sum_{u \in A(s)} \sum_{t \in K(u)} \mathbb{E} \| \nabla \varphi_{u, s} (f(w^{s}_{u_0})) \|^2 \leq \frac{\mathbb{E} [f(w^{s}) - f(w^{s+1})]}{\Gamma_s} \]  
where \( \Gamma_s = \min \{ \Gamma_{u_t} \} \), \( u_0 \) denotes the start iteration during epoch \( u \). Using the update rule of VFB^2-SVRG and summing up all outer loops, and defining \( w_0 \) as initial point and \( w^* \) as optimal solution, we have the final inequality:  
\[ \frac{1}{T} \sum_{s=1}^{S} \sum_{u \in A(s)} \sum_{t \in K(u)} \mathbb{E} \| \nabla f(w^{s}_{u_t}) \|^2 \leq \frac{\mathbb{E} [f(w_0) - f(w^*)]}{TT_s} \]  
where \( T \) denotes the total number of epoches, \( u_0 \) denotes the start iteration during epoch \( u \).  
To prove Theorem 5 set \( \{ c^*_s \}_{u_t = \bar{s}} = 0, \gamma = \frac{m_0}{L_n n^\alpha}, \beta_t = \beta = 2L_n, \) where \( 0 < m_0 < 1, \) and \( 0 < \alpha < 1. \) And there is  
\[ \theta = \gamma \beta = \frac{2m_0}{n^\alpha} \]  
From the recurrence formula of \( c_t \), we have:  
\[ c_0 = \frac{\gamma L_n^2 (1 + \frac{\theta}{n^\alpha})^N - 1}{\theta} \leq \frac{m_0 L_n^2 n^\alpha}{2L_n n^\alpha} \left( (1 + \frac{\theta}{n^\alpha})^N - 1 \right) \leq \frac{L_n}{4} \left( 1 + \frac{\theta}{n^\alpha} - 1 \right) \leq \frac{L_n}{4} (e - 1) \]  
where (a) follow form \( N \leq \left\lfloor \frac{n^\alpha}{2m_0} \right\rfloor \), (b) follows from that \( (1 + \frac{1}{\gamma})^l \) is increasing for \( l > 0, \) and \( \lim_{l \to \infty} (1 + \frac{1}{\gamma})^l = e. \) Since \( e - 1 < 2, \) there is \( c_0 \leq \frac{L_n}{4} \) which satisfies \( c_t \leq \frac{L_n}{4} = L_n \) (used in [138]). Therefore, \( c_t \) is decreasing with respect to \( t, \) and \( c_0 \) is also upper bounded.  
\[ \Gamma_s = \min_l \Gamma_l \geq \frac{\frac{\gamma}{2} - \frac{2}{1-20L^2_{u_t}\gamma^2} (10L^2_{u_t}\gamma^3 - \gamma^2 L_{u_t} + 2c_{u_{t+1}}\gamma^2)}{2 + 4\lambda_{u_t} L_{u_t}^2 L_{u_t}^2} \]  
\[ \geq \frac{\frac{\gamma}{2} - \frac{2n^2\gamma^2}{n^2 - 20m^2\gamma^2} (10m^2\gamma^2 + 2m_0\gamma^2)}{2 + \frac{8L^2_{u_t} m^2}{n^2 - 20m^2\gamma^2}} \]  
\[ \geq \frac{\frac{1}{2} \left( (20m^2_{u_t} + 4\theta) \right) \gamma}{2 + 8L^2_{u_t} m^2_{u_t}} \]  
\[ \geq \frac{\sigma}{L_n n^\alpha} \]
where (a) follows from $c_0 = \max\{c_i\}$, (b) follow form $n^\alpha \leq n^{2\alpha} - 20m_0^2\tau$ (we assume $n \geq \frac{1 + \sqrt{1 + 8m_0^2\tau}}{2}$, this is easy to satisfy when $n$ is large) and $n^\alpha > 1$, (c) follows from that if $\frac{1}{2} > 20m_0^2\tau + 4m_0$ and $\sigma$ is a small value which is independent of $n$.

Above all, if $\tau < \min\{\frac{n^{2\alpha}}{20m_0^2}, \frac{1 - 8m_0}{n^{2\alpha}}\}$ (where $\tau < \frac{n^{2\alpha}}{20m_0^2}$ denotes $\lambda > 0$), where $1 - 8m_0 > 0$, and $N$, satisfies $N \leq \lfloor \frac{n^\alpha}{2m_0} \rfloor$ we have the conclusion:

$$\frac{1}{T} \sum_{s=1}^{S} \sum_{i=0}^{N-1} \mathbb{E}[\|\nabla f(w_i^s)\|^2] \leq \frac{L_n n^\alpha \mathbb{E}[f(w_0) - f(w^*)]}{T\sigma}$$

(154)

where $T$ denotes the total number of epoches. Let R.H.S. of $154 \leq \epsilon$, one can obtain that

$$T \geq \frac{L_n n^\alpha \mathbb{E}[f(w_0) - f(w^*)]}{\epsilon\sigma}$$

(155)

This completes the proof.

**Proof of Theorem[6]**

**Lemma 14.** For all $\forall \psi(t)$, there are

$$\sum_{t=0}^{S-1} \mathbb{E}[\|\hat{\alpha}_t^{\psi(t)}\|^2] \leq \lambda_s \sum_{t=0}^{S-1} \mathbb{E}[\|\hat{\psi}_t^{\psi(t)}\|^2],$$

(156)

where $\lambda_s = \frac{2}{1 - \lambda_0} > 0$.

**Proof of Lemma[14]** First, we give the upper bound to $\mathbb{E}[\|\hat{\psi}_t^{\psi(u)} - \hat{\alpha}_t^{\psi(u)}\|^2]$ as follows. We have that

$$\mathbb{E}[\|\hat{\psi}_t^{\psi(u)} - \hat{\alpha}_t^{\psi(u)}\|^2]$$

(157)

$$= \mathbb{E}[\left(\nabla_{\psi(t)} L(\bar{w}) + \nabla_{\psi(t)} g((\bar{w})_{\psi(v(t))}) - \nabla_{\psi(t)} f(\bar{w}_t) - \hat{\alpha}_t^{\psi(u)} + \hat{\alpha}_t^{\psi(u)} + \frac{1}{n} \sum_{i=1}^{n} \hat{\alpha}_i^{\psi(u)} - \frac{1}{n} \sum_{i=1}^{n} \hat{\psi}_i^{\psi(u)}\right)]^2$$

$$\leq 3\mathbb{E}Q_1 + 3\mathbb{E}Q_2 + 3\mathbb{E}Q_3$$

where $Q_1 = \|\nabla_{\psi(t)} L(\bar{w}) + \nabla_{\psi(t)} g((\bar{w})_{\psi(v(t))}) - \nabla_{\psi(t)} f(\bar{w}_t)\|$ and inequality (a) uses $\|\sum_{i=1}^{n} a_i\|^2 \leq n \sum_{i=1}^{n} \|a_i\|^2$. We will give the upper bounds for the expectations of $Q_1, Q_2$ and $Q_3$ respectively.

$$\mathbb{E}Q_1 = \mathbb{E}[\|\nabla_{\psi(t)} L(\bar{w}) + \nabla_{\psi(t)} g((\bar{w})_{\psi(v(t))}) - \nabla_{\psi(t)} f(\bar{w}_t)\|^2]$$

(158)

above inequality can be obtained by following the proof of Lemma[3]

$$\mathbb{E}Q_2 = \mathbb{E}\left[\|\hat{\alpha}_t^{\psi(t)} - \hat{\hat{\alpha}}_t^{\psi(t)}\|^2\right]$$

(159)

where the inequality uses Lemma[7]

$$\mathbb{E}Q_3 = \mathbb{E}\left[\left\|\frac{1}{n} \sum_{i=1}^{n} \hat{\alpha}_i^{\psi(t)} - \frac{1}{n} \sum_{i=1}^{n} \hat{\alpha}_i^{\psi(t)}\right\|^2\right]$$

(160)

$$\leq \frac{4\tau_2 L_2^2 \gamma^2}{n} \sum_{\phi(t)^{-1}} \sum_{t'=1}^{u \in D'(\xi(t',\psi(t)))} \left(1 - \frac{1}{n}\right)^{\phi(t)^{-1}} \mathbb{E}[\|\hat{\psi}_u^{\psi(t)}\|^2]$$
where the first inequality uses \( \| \sum_{i=1}^{n} a_i \|^2 \leq n \sum_{i=1}^{n} \| a_i \|^2 \), the second inequality uses Lemma 7. Combining (158), (159) and (160) one can obtain:

\[
E \left\| \bar{q}_t^{\psi(t)} - \bar{q}_t^{\psi(t)} \right\|^2 \\
\leq 3E(Q_1 + 3E(Q_2 + 3E(Q_3)) \\
\leq 12L_t^2 \gamma^2 \tau_2 \sum_{t=0}^{S-1} \mathbb{E} \| \bar{v}_t^{\psi(t)} \|^2 + \frac{24\tau_2^2 L_s^2 \gamma^2}{n} \sum_{t'=1}^{S-1} \sum_{u \in D'((\xi(t'),(t')))} \left( 1 - \frac{1}{n} \right)^{\phi(t')-t'-1} \mathbb{E} \| \bar{v}_t^{\psi(u)} \|^2 \\
\] (161)

Summing above equality over all iterations, we have

\[
\sum_{t=0}^{S-1} \mathbb{E} \left\| \bar{v}_t^{\psi(t)} - \bar{v}_t^{\psi(t)} \right\|^2 \\
\leq 12L_t^2 \gamma^2 \tau_2 \sum_{t=0}^{S-1} \mathbb{E} \| \bar{v}_t^{\psi(t)} \|^2 + \frac{24\tau_2^2 L_s^2 \gamma^2}{n} \sum_{t=0}^{S-1} \sum_{t'=1}^{S-1} \left( 1 - \frac{1}{n} \right)^{\phi(t')-t'-1} \mathbb{E} \| \bar{v}_t^{\psi(t')} \|^2 \\
\leq 36L_t^2 \gamma^2 \tau_2 \sum_{t=0}^{S-1} \mathbb{E} \| \bar{v}_t^{\psi(t')} \|^2 \\
\]

(162)

Define \( \bar{v}_t^{\psi(t)} = \nabla \hat{g}_{\psi(t)} f_i(w_t) - \alpha_i^{\psi(t)} + \frac{1}{n} \sum_{i=1}^{n} \alpha_i^{\psi(t)} \). And then, we give the upper bound to \( \mathbb{E} \left\| \bar{v}_t^{\psi(t)} - v_t^{\psi(t)} \right\|^2 \) as follows. We have that

\[
\mathbb{E} \left\| \bar{v}_t^{\psi(t)} - v_t^{\psi(t)} \right\|^2 \\
= \mathbb{E} \left\| \nabla \hat{g}_{\psi(t)} f_i(w_t) - \alpha_i^{\psi(t)} + \frac{1}{n} \sum_{i=1}^{n} \sum_{i=1}^{n} \alpha_i^{\psi(t)} - \nabla \hat{g}_{\psi(t)} f_i(w_t) + \alpha_i^{\psi(t)} - \frac{1}{n} \sum_{i=1}^{n} \alpha_i^{\psi(t)} \right\|^2 \\
\leq \mathbb{E} \left\| \nabla \hat{g}_{\psi(t)} f_i(w_t) - \nabla \hat{g}_{\psi(t)} f_i(w_t) \right\|^2 + 3\mathbb{E} \left\| \alpha_i^{\psi(t)} - \alpha_i^{\psi(t)} \right\|^2 + 3\mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \alpha_i^{\psi(t)} - \frac{1}{n} \sum_{i=1}^{n} \alpha_i^{\psi(t)} \right\|^2 \\
\]

where the inequality (a) uses \( \| \sum_{i=1}^{n} a_i \|^2 \leq n \sum_{i=1}^{n} \| a_i \|^2 \). We will give the upper bounds for the expectations of \( Q_4, Q_5 \) and \( Q_6 \) respectively.

\[
\mathbb{E}Q_4 = \mathbb{E} \left\| \nabla \hat{g}_{\psi(t)} f_i(w_t) - \nabla \hat{g}_{\psi(t)} f_i(w_t) \right\|^2 \\
\leq L^2_{\psi(t)} \mathbb{E} \left\| w_t - \hat{w}_t \right\|^2 \\
\leq L^2_{\psi(t)} \gamma^2 \mathbb{E} \left\| \sum_{t' \in D(u,t)} U_{\psi(t')} v_t^{\psi(t')} \right\|^2 \\
\leq \tau_1 L^2_{\gamma^2} \gamma^2 \sum_{t' \in D(u,t)} \mathbb{E} \left\| v_t^{\psi(t')} \right\|^2 \\
\]

where (a) uses Assumption 2, (b) uses \( \| \sum_{i=1}^{n} a_i \|^2 \leq n \sum_{i=1}^{n} \| a_i \|^2 \). Similar to the analyses of \( Q_2 \) and \( Q_3 \), we have

\[
\mathbb{E}Q_5 = \mathbb{E} \left\| \alpha_i^{\psi(t)} - \alpha_i^{\psi(t)} \right\|^2 \\
\leq \frac{\tau_1 L^2_{\gamma^2} \gamma^2}{n} \sum_{t'=1}^{\phi(t')-t'-1} \sum_{\hat{u} \in D((\xi(t'),(t')))} \left( 1 - \frac{1}{n} \right)^{\phi(t')-t'-1} \mathbb{E} \left\| \bar{v}_t^{\psi(u)} \right\|^2 \\
\]

(165)

where the inequality uses Lemma 7.

\[
\mathbb{E}Q_6 = \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \alpha_i^{\psi(t)} - \frac{1}{n} \sum_{i=1}^{n} \alpha_i^{\psi(t)} \right\|^2 \\
\leq \frac{\tau_1 L^2_{\gamma^2} \gamma^2}{n} \left( 1 - \frac{1}{n} \right)^{\phi(t')-t'-1} \sum_{\hat{u} \in D((\xi(t'),(t')))} \mathbb{E} \left\| \bar{v}_t^{\psi(u)} \right\|^2 \\
\]
Summing above inequality for \( t = 0, \ldots, S - 1 \) and follow the analyses of Eq. (163) one can have

\[
\sum_{t=0}^{S-1} \mathbb{E} \left\| \tilde{v}_u^{\psi(t)} - \tilde{v}_t^{\psi(t)} \right\|^2 \\
\leq 3L_*^2 \gamma^2 \tau_t^2 \sum_{t=0}^{S-1} \mathbb{E} \left\| \tilde{v}_t^{\psi(t')} \right\|^2 + \frac{6\tau_t^2 L_*^2 \gamma^2}{n} \sum_{t=0}^{S-1} \phi(t)-1 \sum_{t'=1}^{S-1} \left(1 - \frac{1}{n}\right) \phi(t') \mathbb{E} \left\| \tilde{v}_u^{\psi(\hat{u})} \right\|^2 \\
\leq 9L_*^2 \gamma^2 \tau_t^2 \sum_{t=0}^{S-1} \mathbb{E} \left\| \tilde{v}_t^{\psi(t')} \right\|^2
\]

Based on above formulations, we have

\[
\sum_{t=0}^{S-1} \mathbb{E} \left\| \tilde{v}_u^{\psi(t)} - v_t^{\psi(t)} \right\|^2 = \sum_{t=0}^{S-1} \mathbb{E} \left\| \tilde{v}_u^{\psi(t)} - \tilde{v}_u^{\psi(t)} + \tilde{v}_u^{\psi(t)} - v_t^{\psi(t)} \right\|^2 \\
\leq \sum_{t=0}^{S-1} \left( 2\mathbb{E} \left\| \tilde{v}_u^{\psi(t)} - \tilde{v}_u^{\psi(t)} \right\|^2 + 2\mathbb{E} \left\| \tilde{v}_u^{\psi(t)} - v_t^{\psi(t)} \right\|^2 \right) \\
\leq 18L_*^2 \gamma^2 \tau_t^2 \sum_{t=0}^{S-1} \mathbb{E} \left\| \tilde{v}_t^{\psi(t')} \right\|^2 + 72L_*^2 \gamma^2 \tau_t^2 \sum_{t=0}^{S-1} \mathbb{E} \left\| \tilde{v}_t^{\psi(t)} \right\|^2
\]

(167)

then we have

\[
\sum_{t=0}^{S-1} \mathbb{E} \left\| \tilde{v}_t^{\psi(t)} \right\|^2 = \sum_{t=0}^{S-1} \mathbb{E} \left\| \tilde{v}_t^{\psi(t)} - v_t^{\psi(t)} + v_t^{\psi(t)} \right\|^2 \leq \sum_{t=0}^{S-1} \left( 2\mathbb{E} \left\| \tilde{v}_t^{\psi(t)} - v_t^{\psi(t)} \right\|^2 + 2\mathbb{E} \left\| v_t^{\psi(t)} \right\|^2 \right) \\
\leq L_*^2 \gamma^2 \left( 36\tau_t^2 + 144\tau_t^2 \right) \sum_{t=0}^{S-1} \mathbb{E} \left\| \tilde{v}_t^{\psi(t')} \right\|^2 + 2 \sum_{t=0}^{S-1} v_t^{\psi(t)}
\]

(168)

which implies that if \( \lambda_\gamma = \frac{2}{1 - 180 \eta_\tau L_*^2 \gamma^2 \tau_t^2} > 0 \), we have

\[
\sum_{t=1}^{S} \mathbb{E} \left\| \tilde{v}_t^{\psi(t)} \right\|^2 \leq \lambda_\gamma \sum_{t=1}^{S} \mathbb{E} \left\| v_t^{\psi(t)} \right\|^2
\]

(169)

This completes the proof. \( \square \)

Similar to the proof of Theorem 4, we first apply Lemma 4 to all \( S \) iterations and there is

\[
\sum_{t \in A(S)} \mathbb{E} \left\| \nabla f(w_t) \right\|^2 \leq 2L_*^2 \gamma^2 \eta_t^2 \sum_{t=0}^{S-1} \mathbb{E} \left\| \tilde{v}_t^{\psi(t)} \right\|^2 + 2 \sum_{t=0}^{S-1} \mathbb{E} \left\| \nabla g_{\psi(t)} f(w_t) \right\|^2
\]

(170)

Then, we give the upper bound to \( \mathbb{E} \left\| v_t^{\psi(t)} \right\|^2 \) as follows. We define:

\[
\zeta_t^{\psi(t)} = \nabla g_{\psi(t)} f_t(w_t) - \alpha_t^{t,\psi(t)}
\]

(171)

and use the definition of \( v_t^{\psi(t)} \) to get

\[
\mathbb{E} \left\| v_t^{\psi(t)} \right\|^2 = \mathbb{E} \left\| \zeta_t^{\psi(t)} + \frac{1}{n} \sum_{i=1}^{n} \alpha_i^{t,\psi(t)} \right\|^2 \\
= \mathbb{E} \left\| \zeta_t^{\psi(t)} + \frac{1}{n} \sum_{i=1}^{n} \alpha_i^{t,\psi(t)} - \nabla g_{\psi(t)} f(w_t) + \nabla g_{\psi(t)} f(w_t) \right\|^2 \\
\leq 2\mathbb{E} \left\| \nabla g_{\psi(t)} f(w_t) \right\|^2 + 2\mathbb{E} \left\| \zeta_t^{\psi(t)} - \mathbb{E} \left[ \zeta_t^{\psi(t)} \right] \right\|^2 \\
\leq 2\mathbb{E} \left\| \nabla g_{\psi(t)} f(w_t) \right\|^2 + 2\mathbb{E} \left\| \zeta_t^{\psi(t)} \right\|^2
\]

(172)
where (a) follows from $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ and $E\left[\psi_i^t\right] = \nabla g_{\psi(t)} f(w_t) - \frac{1}{n} \sum_{i=1}^n \alpha_i^{t,\psi(t)}$, (b) follows from $E\left[\psi_i^t\right] - E\left[\psi_i^t\right]^2 \leq E\|\psi_i^t\|^2$, we have

$$E\|\psi_i^t\|^2$$

\[ \leq 2E\|\nabla g_{\psi(t)} f(w_t)\|^2 + 2E\|\psi_i^t - \nabla g_{\psi(t)} f_i(w_t)\|^2 \]

\[ \leq 2E\|\nabla g_{\psi(t)} f(w_t)\|^2 + 2E\|\psi_i^t - \nabla g_{\psi(t)} f_i(w_t)\|^2 \]

\[ \leq 2E\|\nabla g_{\psi(t)} f(w_t)\|^2 + 2L^2 \left( 1 - \frac{1}{n} \right)^{\phi(t)-1} E\|w_0 - w_t\|^2 \]

\[ + 2L^2 \frac{1}{n} \sum_{t'=1}^{\phi(t)-1} \left( 1 - \frac{1}{n} \right)^{\phi(t)-t'-1} E\|w_{\xi(t',\psi(t))} - w_t\|^2 \]

\[ \leq 2E\|\nabla g_{\psi(t)} f(w_t)\|^2 + 2L^2 \left( 1 - \frac{1}{n} \right)^{\phi(t)-1} E\|w_0 - w_t\|^2 \]

\[ + 2L^2 \frac{1}{n} \sum_{t'=1}^{\phi(t)-1} \left( 1 - \frac{1}{n} \right)^{\phi(t)-t'-1} \left( 2E\|w_{\xi(t',\psi(t))} - w_0\|^2 + 2E\|w_0 - w_t\|^2 \right) \]

\[ \leq 2E\|\nabla g_{\psi(t)} f(w_t)\|^2 + 6L^2 E\|w_0 - w_t\|^2 + 4L^2 \frac{1}{n} \sum_{t'=1}^{\phi(t)-1} \left( 1 - \frac{1}{n} \right)^{\phi(t)-t'-1} E\|w_{\xi(t',\psi(t))} - w_0\|^2 \] (173)

where (a) follows from Lemma 7, (b) follows from Assumption 2, (c) follows from $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, (d) follows from $\sum_{t'=1}^{\phi(t)-1} \left( 1 - \frac{1}{n} \right)^{\phi(t)-t'-1} < 1$ and $\left( 1 - \frac{1}{n} \right)^{\phi(t)-1} < 1$. Moreover, there is

$$\sum_{t=0}^{S-1} \frac{1}{n} \sum_{t'=1}^{\phi(t)-1} \left( 1 - \frac{1}{n} \right)^{\phi(t)-t'-1} E\|w_{\xi(t',\psi(t))} - w_0\|^2 \leq \sum_{t=0}^{S-1} E\|w_t - w_0\|^2$$ (174)

which follows from that $\frac{1}{n} \sum_{t'=1}^{\phi(t)-1} \left( 1 - \frac{1}{n} \right)^{\phi(t)-t'-1} \leq 1$. As for $E\|w_0 - w_t\|^2$ there is

$$E\|w_0 - w_t\|^2 = E\|w_0 - w_0 + w_t - w_t\|^2$$

$$= E\|w_0 - w_t\|^2 + E\|w_t - w_{t+1}\|^2 - 2E\langle w_0 - w_t, w_t - w_{t+1} \rangle$$

$$= E\|w_0 - w_t\|^2 + E\|w_t - w_{t+1}\|^2 - 2\gamma E\langle w_0 - w_t, \nabla g_{\psi(t)} \rangle$$

$$\leq E\|w_0 - w_t\|^2 + E\|w_t - w_{t+1}\|^2 + 2\gamma \left( \frac{1}{2\beta_t} E\|\nabla g_{\psi(t)}\|^2 + \frac{\beta_t}{2} E\|w_0 - w_t\|^2 \right)$$

$$= (1 + \gamma \beta_t) E\|w_0 - w_t\|^2 + \gamma^2 E\|w_t - w_{t+1}\|^2 + \frac{\gamma}{\beta_t} E\|\nabla g_{\psi(t)}\|^2$$ (175)

**Proof of Theorem 6** First, we upper bound $E f(w_{t+1})$ for $t = 0, \ldots, S - 1$:

$$E f(w_{t+1}) \leq E \left[ f(w_{t+1}) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{L}{2} \|w_{t+1} - w_t\|^2 \right]$$

$$= E f(w_t) - \gamma E \left\langle \nabla f(w_t), \nabla g_{\psi(t)} \right\rangle + \frac{\gamma^2 L}{2} E\|\psi(t)\|^2$$

$$= \left( b \right) E f(w_t) - \gamma E \left[ \|\nabla g_{\psi(t)} f(w_t)\|^2 + \|\nabla g_{\psi(t)}\|^2 \right]$$

$$\text{Proof of Theorem 6}$$

First, we upper bound $E f(w_{t+1})$ for $t = 0, \ldots, S - 1$:

$$E f(w_{t+1}) \leq E \left[ f(w_{t+1}) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{L}{2} \|w_{t+1} - w_t\|^2 \right]$$

$$= E f(w_t) - \gamma E \left\langle \nabla f(w_t), \nabla g_{\psi(t)} \right\rangle + \frac{\gamma^2 L}{2} E\|\psi(t)\|^2$$

$$\text{Proof of Theorem 6}$$

First, we upper bound $E f(w_{t+1})$ for $t = 0, \ldots, S - 1$:

$$E f(w_{t+1}) \leq E \left[ f(w_{t+1}) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{L}{2} \|w_{t+1} - w_t\|^2 \right]$$

$$= E f(w_t) - \gamma E \left\langle \nabla f(w_t), \nabla g_{\psi(t)} \right\rangle + \frac{\gamma^2 L}{2} E\|\psi(t)\|^2$$

$$\text{Proof of Theorem 6}$$

First, we upper bound $E f(w_{t+1})$ for $t = 0, \ldots, S - 1$:

$$E f(w_{t+1}) \leq E \left[ f(w_{t+1}) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{L}{2} \|w_{t+1} - w_t\|^2 \right]$$

$$= E f(w_t) - \gamma E \left\langle \nabla f(w_t), \nabla g_{\psi(t)} \right\rangle + \frac{\gamma^2 L}{2} E\|\psi(t)\|^2$$

$$\text{Proof of Theorem 6}$$

First, we upper bound $E f(w_{t+1})$ for $t = 0, \ldots, S - 1$:

$$E f(w_{t+1}) \leq E \left[ f(w_{t+1}) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{L}{2} \|w_{t+1} - w_t\|^2 \right]$$

$$= E f(w_t) - \gamma E \left\langle \nabla f(w_t), \nabla g_{\psi(t)} \right\rangle + \frac{\gamma^2 L}{2} E\|\psi(t)\|^2$$
where (a) follows from Assumption 2, (b) follows form \( \langle a, b \rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2 \). Next, we give the upper bound of the term \( \mathbb{E} |\nabla g_{v(e)} f(w_t) - \nabla g_{v(e)}|^2 |^2 \):

\[
\mathbb{E} |\nabla g_{v(e)} f(w_t) - \nabla g_{v(e)}|^2 |^2 \leq 2L^2_\gamma^2 \gamma^2 \tau_1 \sum_{u \in D'(t)} \mathbb{E} \|\tilde{v}_{u}^{(t')} |^2 + 8L^2_\gamma^2 \gamma^2 \tau_2 \sum_{u \in D'(t)} \mathbb{E} \|\tilde{v}_{u}^{(t')} |^2.
\] (177)

Above result can be obtained by following the analyses of Lemma 3. From Eqs. (176) and (177), it is easy to derive the following inequality:

\[
\mathbb{E} f(w_{t+1}) \leq \mathbb{E} f(w_t) - \gamma \mathbb{E} |\nabla g_{v(e)} f(w_t)|^2 - \gamma \mathbb{E} |\nabla g_{v(e)}|^2 + \frac{\gamma^2 L}{2} \mathbb{E} |\tilde{v}_t^{(t')} |^2 + L^2_\gamma^3 \left( \tau_1 \sum_{u \in D(u)} \mathbb{E} \|\tilde{v}^{(t')} |^2 + 4 \tau_2 \sum_{u \in D'(u)} \mathbb{E} \|\tilde{v}^{(t')} |^2 \right)
\]

(178)

Here, we define a Lyapunov function:

\[
R_t = \mathbb{E} f(w_t) + c_t \mathbb{E} |w_0 - w_t|^2.
\] (179)

From the definition of Lyapunov function, and (179):

\[
R_{t+1} \leq \mathbb{E} f(w_{t+1}) + c_{t+1} \mathbb{E} |w_0 - w_{t+1}|^2
\]

\[
\leq \mathbb{E} f(w_t) - \gamma \mathbb{E} |\nabla g_{v(e)} f(w_t)|^2 - \gamma \mathbb{E} |\nabla g_{v(e)}|^2 + \frac{\gamma^2 L}{2} \mathbb{E} |\tilde{v}_t^{(t')} |^2 + L^2_\gamma^3 \left( \tau_1 \sum_{u \in D(u)} \mathbb{E} \|\tilde{v}^{(t')} |^2 + 4 \tau_2 \sum_{u \in D'(u)} \mathbb{E} \|\tilde{v}^{(t')} |^2 \right) + c_{t+1} \mathbb{E} |w_0 - w_{t+1}|^2
\]

\[
\leq \mathbb{E} f(w_t) - \gamma \mathbb{E} |\nabla g_{v(e)} f(w_t)|^2 - \gamma \mathbb{E} |\nabla g_{v(e)}|^2 + \frac{\gamma^2 L}{2} \mathbb{E} |\tilde{v}_t^{(t')} |^2 + L^2_\gamma^3 \left( \tau_1 \sum_{u \in D(u)} \mathbb{E} \|\tilde{v}^{(t')} |^2 + 4 \tau_2 \sum_{u \in D'(u)} \mathbb{E} \|\tilde{v}^{(t')} |^2 \right)
\]

\[
+ c_{t+1} \left( (1 + \gamma \beta_t) \mathbb{E} |w_0 - w_t|^2 + \frac{\gamma}{\beta_t} \mathbb{E} |\nabla g_{v(e)}|^2 + \gamma^2 \mathbb{E} |\tilde{v}_t^{(t')} |^2 \right)
\]

(180)

where (a) follows from Eq. (175). Summing above inequality for all iterations then we have that:

\[
\sum_{t=0}^{S-1} R_{t+1} = \sum_{t=0}^{S-1} \left( \mathbb{E} f(w_t) - \gamma \mathbb{E} |\nabla g_{v(e)} f(w_t)|^2 - (\gamma - \frac{\gamma c_{t+1}}{\beta_t}) \mathbb{E} |\nabla g_{v(e)}|^2 \right)
\]

\[
+ \sum_{t=0}^{S-1} \left( \frac{L^2_\gamma^3 \tau_1^2}{2} + 4 \frac{\gamma^2 L}{2} \mathbb{E} |\tilde{v}_t^{(t')} |^2 \right)
\]

\[
+ c_{t+1} \mathbb{E} |w_0 - w_{t+1}|^2
\]

\[
\leq \mathbb{E} f(w_t) - \gamma \mathbb{E} |\nabla g_{v(e)} f(w_t)|^2 - (\gamma - \frac{\gamma c_{t+1}}{\beta_t}) \mathbb{E} |\nabla g_{v(e)}|^2
\]

\[
+ \sum_{t=0}^{S-1} \left( \frac{5L^2_\gamma^3 \tau_1^2}{2} + c_{t+1} \gamma \beta_t \right) \mathbb{E} |\tilde{v}_t^{(t')} |^2 + c_{t+1} \left( (1 + \gamma \beta_t) \mathbb{E} |w_0 - w_t|^2 \right)
\]

\[
\leq \sum_{t=0}^{S-1} \left( \mathbb{E} f(w_t) - \frac{\gamma}{2} \mathbb{E} |\nabla g_{v(e)} f(w_t)|^2 - (\frac{\gamma}{2} - \frac{\gamma c_{t+1}}{\beta_t}) \mathbb{E} |\nabla g_{v(e)}|^2 \right)
\]

\[
+ \sum_{t=0}^{S-1} \left( \frac{5L^2_\gamma^3 \tau_1^2}{2} + c_{t+1} (1 + \gamma \beta_t) \mathbb{E} |w_0 - w_t|^2 \right)
\]

\[
+ \sum_{t=1}^{S} \lambda_t \mathbb{E} |\nabla g_{v(e)} f(w_t)|^2 + 5L^2_\gamma |w_0 - w_t|^2
\]

\[
+ \sum_{t=0}^{S-1} c_{t+1} \left( (1 + \gamma \beta_t) \mathbb{E} |w_0 - w_t|^2 \right)
\]
\[
\begin{align*}
\sum_{t=0}^{S-1} (E(f(w_t)) + (c_{t+1}(1 + \gamma\beta) + 5L^2 \lambda_t \lambda_t) \|w_0 - w_t\|^2)
&= \sum_{t=0}^{S-1} (\frac{\gamma}{2} - \lambda_t \lambda_t)E\|\nabla g_{u(t)} f (w_t)\|^2 - \sum_{t=0}^{S-1} (\frac{\gamma}{2} - \frac{\gamma\lambda_{t+1}}{\beta_t})E\|\nabla g_{u(t)}\|^2
\end{align*}
\]
where (a) follows from the definition of \(\tau\), (b) uses Eqs. [173] and [174], \(\lambda_t = 10L^2 \gamma^3 \tau + \gamma^2 L + 2c_{t+1}\gamma^2\), and the definition of \(L_\ast\).

Similarly to the proof of Theorem 5, we have
\[
\sum_{u \in \mathcal{A}(S)} \sum_{t \in K'(u)} E\|\nabla g_{u(t)} f (w_{u_0})\|^2
\]
where \(u_0\) denotes the start global iteration of epoch \(u\), (a) follows from Eq. [170] (b) follows from Eqs. [173] and [174]. This implies that
\[
\frac{\gamma}{2} - \lambda_t \lambda_t \sum_{u \in \mathcal{A}(S)} \sum_{t \in K'(u)} E\|\nabla g_{u(t)} f (w_{u_0})\|^2
\]
where (a) follows from the definition of \(L_\ast\). Combining Eq. [183] with [181] we have
\[
\left(\frac{\gamma}{2} - \lambda_t \lambda_t \right) \sum_{u \in \mathcal{A}(S)} \sum_{t \in K'(u)} \left( E(f(w_t)) + (c_{t+1}(1 + \gamma\beta) + 5L^2 \lambda_t \lambda_t) E\|w_0 - w_t\|^2 \right)
- \sum_{u \in \mathcal{A}(S)} \sum_{t \in K'(u)} R_{t+1}
\end{align*}
\]
Rearrange Eq. [184] we have
\[
\sum_{u \in \mathcal{A}(S)} \sum_{t \in K'(u)} R_{t+1} \leq \sum_{u \in \mathcal{A}(S)} \sum_{t \in K'(u)} \left( E(f(w_t)) + (c_{t+1}(1 + \gamma\beta) + \frac{5}{2}L^2 \gamma^2) E\|w_0 - w_t\|^2 \right)
- \frac{\gamma}{2} - \lambda_t \lambda_t \sum_{u \in \mathcal{A}(S)} \sum_{t \in K'(u)} E\|\nabla g_{u(t)} f (w_{u_0})\|^2
\]
where
\[
c_t = c_{t+1}(1 + \gamma\beta_t) + \frac{5}{2}L^2 \gamma^2
\]
and

\[
\Gamma_t = \frac{\frac{\gamma^2}{2} - \frac{2}{1-180L_\star^2\gamma^2\tau}(10L_\star^2\gamma^3\tau + \gamma^2L_\star + 2c_0\gamma^2)}{2 + 4\lambda_\gamma L_\star^4\gamma^2\tau} \tag{187}
\]

Let \(\bar{S}\) be the subscript of the final global iteration, and one can set \(\{c_t\}_{t=S} = 0\), define \(w_0\) as initial point and \(w^*\) as optimal solution, we have

\[
\frac{1}{\bar{S}} \sum_{u \in A(\bar{S})} \sum_{t \in K'(u)} \Gamma_t \mathbb{E}\|\nabla \psi(t)f(w_u)\|^2 \leq \frac{\mathbb{E}[f(w_0) - f(w^*)]}{\bar{S} \Gamma_*} \tag{188}
\]

where \(\bar{t}\) denotes the start global iteration of epoch \(t\) and use \(\Gamma_* = \min_t \{\Gamma_t\}\).

To prove Theorem 6, set \(\{c_t\}_{t=\bar{S}-1} = 0\), \(\gamma = \frac{m_0}{L_\star n^{\alpha}}, \beta_t = \beta = 4L_\star\), where \(0 < m_0 < 1\), and \(0 < \alpha < 1\). And there is

\[
\theta = \gamma \beta_t = \frac{4m_0}{n^{\alpha}} \tag{189}
\]

Then following the analysis of Eq. 152, we have that the total epoch number \(T\) should satisfy

\[
T \geq \left\lfloor \frac{n^{\alpha}}{4m_0} \right\rfloor
\]

Based on above analyses, we have the conclusion:

\[
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(w_t)\|^2 \leq \frac{L_\star n^{\alpha} \mathbb{E}[f(w_0) - f(w^*)]}{T \sigma} \tag{191}
\]

where, \(T\) denotes the number of total epoches, \(t_0\) is the start iteration of epoch \(t\). This completes the proof.