Momentum Benefits Non-IID Federated Learning Simply and Provably

Ziheng Cheng*[†] Xinmeng Huang*[‡] Kun Yuan[§][¶]

Abstract

Federated learning is a powerful paradigm for large-scale machine learning, but it faces significant challenges due to unreliable network connections, slow communication, and substantial data heterogeneity across clients. FedAvg and SCAFFOLD are two fundamental algorithms to address these challenges. In particular, FedAvg employs multiple local updates before communicating with a central server, while SCAFFOLD maintains a control variable on each client to compensate for "client drift" in its local updates. Various methods have been proposed in literature to enhance the convergence of these two algorithms, but they either make impractical adjustments to algorithmic structure, or rely on the assumption of bounded data heterogeneity.

This paper explores the utilization of momentum to enhance the performance of FedAvg and SCAFFOLD. When all clients participate in the training process, we demonstrate that incorporating momentum allows FedAvg to converge without relying on the assumption of bounded data heterogeneity even using a constant local learning rate. This is a novel result since existing analyses for FedAvg require bounded data heterogeneity even with diminishing local learning rates. In the case of partial client participation, we show that momentum enables SCAFFOLD to converge provably faster without imposing any additional assumptions. Furthermore, we use momentum to develop new variance-reduced extensions of FedAvg and SCAFFOLD, which exhibit state-of-the-art convergence rates. Our experimental results support all theoretical findings.

1 Introduction

Federated learning (FL) is a powerful paradigm for large-scale machine learning (Konečnỳ et al., 2016; McMahan et al., 2017a). In situations where data and computational resources are dispersed among a diverse range of clients, including phones, tablets, sensors, hospitals, and other devices and agents, federated learning

*Equal Contribution.

[†] Peking University. alex-czh@stu.pku.edu.cn.

[‡] University of Pennsylvania. xinmengh@sas.upenn.edu.

[§] Peking University. kunyuan@pku.edu.cn.

[¶] Corresponding Author.

facilitates local data processing and collaboration among these clients (Kairouz et al., 2021). Consequently, a centralized model can be trained without transmitting decentralized data from clients directly to servers, thereby ensuring a fundamental level of privacy.

Federated learning encounters several significant challenges in algorithmic development. Firstly, the reliability and relatively slow nature of network connections between the server and clients pose obstacles to efficient communication during the training process. Secondly, the dynamic availability of only a small subset of clients for training at any given time demands strategies that can adapt to this variable environment. Lastly, the presence of substantial heterogeneity of non-iid data across different clients further complicates the training process.

FedAvg (Konečný et al., 2016; McMahan et al., 2017a; Stich, 2019; Yu et al., 2019a; Lin et al., 2020; Wang & Joshi, 2021) has emerged as a prevalent algorithm for federated learning, leveraging multiple stochastic gradient descent (SGD) steps within each client before communicating with a central server. While FedAvg is readily implementable and has demonstrated success in certain applications, its performance is notably hindered by the presence of data heterogeneity, *i.e.*, non-iid clients, even when *all clients* participate in the training process (Li et al., 2019; Yang et al., 2021). To mitigate the influence of data heterogeneity, SCAFFOLD (Karimireddy et al., 2020b) maintains a control variable on each client to compensate for "client drift" in its local SGD updates, making convergence more robust to data heterogeneity and client sampling. Due to their practicality and effectiveness, FedAvg and SCAFFOLD have become foundational algorithms in federated learning, leading to the development of numerous variants that cater to decentralized (Koloskova et al., 2020; Rizk et al., 2022; Nguyen et al., 2022; Alghunaim, 2023), compressed (Haddadpour et al., 2021; Reisizadeh et al., 2020; Mitra et al., 2021), asynchronous (Chen et al., 2020a,b; Xu et al., 2021a), and personalized (Fallah et al., 2020; Pillutla et al., 2022; Tan et al., 2022; T Dinh et al., 2020) federated learning scenarios.

Various methods have been proposed to enhance the convergence of FedAvg, SCAFFOLD, and their variance-reduced[1] extensions. While exhibiting superior convergence rates, these approaches typically make impractical adjustments to algorithmic structures. For instance, STEM (Khanduri et al., 2021) requires increasing either the batch size or the number of local steps with algorithmic iterations. Similarly, CE-LSGD (Patel et al., 2022) and MIME (Karimireddy et al., 2020a) mandate computing a large-batch or even full-batch local gradient per round for each client. Additionally, FedProx (Li et al., 2020), FedPD (Zhang et al., 2021), and FedDyn (Durmus et al., 2021) rely on solving "local problems" to an extremely high precision. These adjustments may not align with the practical constraints in federated learning setups.

Furthermore, many of these algorithms, including FedAvg, STEM, FedProx, MIME, and CE-SGD, still rely on the assumption of bounded data heterogeneity. When this assumption is violated, the theoretical analyses of these algorithms become invalid. While some algorithms, such as LED (Alghunaim, 2023) and

[1] Throughout the paper, variance reduction refers to techniques aiming to mitigate the influence of within-client gradient stochasticity, as opposed to the inter-client data heterogeneity.

Table 1: The comparison of convergence rates of FL algorithms when **all clients** participate in training. Notation L is the smoothness constant of objective functions, $\Delta = f(x^0) - \min_x f(x)$ is the initialization gap, σ^2 is the variance of gradient noises, N is the number of clients, K is the number of local steps per round, and R is the number of communication rounds, ζ^2 and G are uniform bounds of data heterogeneity $(1/N) \sum_{1 \le i \le N} \|\nabla f_i(x) - \nabla f(x)\|^2$ and gradient norm $\max_{1 \le i \le N} \|\nabla f_i(x)\|$ with $G^2 \gg \zeta^2$ typically. The "Assumptions" column lists all assumptions beyond Assumption 1 and 3.

Algorithm	**Convergence Rate** $\mathbb{E}[\|\nabla f(\hat{x})\|^2] \lesssim$	**Assumptions**
FedAvg		
(Yu et al., 2019b)	$\left(\frac{L\Delta\sigma^2}{NKR}\right)^{1/2} + \left(\frac{L\Delta G}{R}\right)^{2/3} + \frac{L\Delta}{R}$	Bounded grad.
(Koloskova et al., 2020)	$\left(\frac{L\Delta\sigma^2}{NR}\right)^{1/2} + \left(\frac{L\Delta K\zeta}{R}\right)^{2/3} + \frac{L\Delta K}{R}$	Bounded hetero.
(Karimireddy et al., 2020b)	$\left(\frac{L\Delta\sigma^2}{NKR}\right)^{1/2} + \left(\frac{L\Delta\zeta}{R}\right)^{2/3} + \frac{L\Delta}{R}$	Bounded hetero.
(Yang et al., 2021)	$\left(\frac{L\Delta\sigma^2}{NKR}\right)^{1/2} + \frac{L\Delta}{R}$	Bounded hetero.[1]
LED (Alghunaim, 2023)	$\left(\frac{L\Delta\sigma^2}{NKR}\right)^{1/2} + \left(\frac{L\Delta\sigma}{\sqrt{K}R}\right)^{2/3} + \frac{L\Delta}{R}$	–
VRL-SGD (Liang et al., 2019)	$\left(\frac{L\Delta\sigma^2}{NKR}\right)^{1/2} + \left(\frac{L\Delta\sigma}{\sqrt{K}R}\right)^{2/3} + \frac{L\Delta}{R}$	–
FedAvg-M (Thm. 1)	$\left(\frac{L\Delta\sigma^2}{NKR}\right)^{1/2} + \frac{L\Delta}{R}$	–
Variance-Reduction		
BVR-L-SGD (Murata & Suzuki, 2021)	$\left(\frac{L\Delta\sigma}{NKR}\right)^{2/3} + \frac{\sigma^2}{NKR} + \frac{L\Delta}{R}$	Sample smooth $\mathcal{O}(K)$ batchsize[2]
CE-LSGD (Patel et al., 2022)	$\left(\frac{L\Delta\sigma}{NKR}\right)^{2/3} + \frac{\sigma^2}{NKR} + \frac{L\Delta}{R}$	Sample smooth $\mathcal{O}(K)$ batchsize[2]
STEM (Khanduri et al., 2021)	$\frac{L\Delta + \sigma^2 + \zeta^2}{(NKR)^{2/3}} + \frac{L\Delta}{R}$	Sample smooth Bounded hetero.
FedAvg-M-VR (Thm. 2)	$\left(\frac{L\Delta\sigma}{NKR}\right)^{2/3} + \frac{L\Delta}{R}$	Sample smooth

[1] The local learning rate vanishes to zero when data heterogeneity is unbounded, *i.e.*, $\zeta \to \infty$.

[2] A large batch is needed by each client per communication round.

VRL-SGD (Liang et al., 2019), can handle unbounded data heterogeneity, their convergence rates are not state-of-the-art, as demonstrated in Table 1. These limitations motivate us to develop novel strategies that are easy to implement, robust to data heterogeneity, and exhibit superior theoretical convergence rates.

1.1 Main results and contributions

This paper examines the utilization of *momentum* to enhance the performance of FedAvg and SCAFFOLD.

In order to ensure simplicity and practicality in implementations, we only introduce momentum to the local SGD steps, avoiding any inclusion of impractical elements, such as gradient computation of large batchsizes or solving local problems to high precision. Remarkably, this straightforward approach effectively alleviates the necessity for stringent assumptions of bounded data heterogeneity, leading to noteworthy improvements in convergence rates. The main findings and contributions of this paper are summarized below.

First, when all clients participate in the training process:

- We demonstrate that incorporating momentum allows FedAvg and its variance-reduced extension to *converge without relying on the assumption of bounded data heterogeneity*, even using constant local learning rates. This is rather surprising as, to our knowledge, all existing analyses for FedAvg, *e.g.*, (Karimireddy et al., 2020b; Yang et al., 2021; Wang et al., 2020b), require bounded data heterogeneity even with diminishing local learning rates.
- We further establish that, by effectively removing the influence of data heterogeneity on convergence, momentum empowers FedAvg and its variance-reduced extension with state-of-the-art convergence rates in the context of full client participation.

Second, when partial clients participate in the training process per iteration:

- The proposed SCAFFOLD-M that incorporates momentum into SCAFFOLD achieves a provably faster convergence rate. To our knowledge, this is the *first* result that improves upon SCAFFOLD without imposing any additional assumptions beyond those used in (Karimireddy et al., 2020b).
- We further introduce momentum to SCAFFOLD with variance reduction, achieving the *first* variance-reduced federated learning algorithm that does not rely on the assumption of bounded data heterogeneity. This algorithm attains a state-of-the-art convergence rate in the context of partial client participation and unbounded data heterogeneity.

Tables 1 and 2 present a comprehensive comparison of the convergence rates and associated assumptions of existing algorithms, as well as our newly proposed approaches.

It is observed that by simply adding momentum to local steps, FedAvg, SCAFFOLD, and their variance-reduced extensions all attain state-of-the-art convergence rates without resorting to further assumptions such as bounded data heterogeneity. We support our theoretical findings with extensive numerical experiments.

Table 2: The comparison of convergence rates of FL algorithms when ***S* clients among *N* ones** participate in training per iteration. Notations are the same as those in Table 1.

Algorithm	**Convergence Rate** $\mathbb{E}[\Vert\nabla f(\hat{x})\Vert^2] \lesssim$	**Assumptions**
SCAFFOLD (Karimireddy et al., 2020b)	$\left(\frac{L\Delta\sigma^2}{SKR}\right)^{1/2} + \frac{L\Delta}{R}\left(\frac{N}{S}\right)^{2/3}$	–
SCAFFOLD-M (Thm. 3)	$\left(\frac{L\Delta\sigma^2}{SKR}\right)^{1/2} + \frac{L\Delta}{R}\left(1+\frac{N^{2/3}}{S}\right)$	–
VARIANCE-REDUCTION		
MIMELITEMVR[1] (Karimireddy et al., 2020a)	$\left(\frac{L\Delta(\sigma+\zeta)}{R}\right)^{2/3} + \frac{L\Delta+\sigma^2+\zeta^2}{R}$	Sample smooth Noiseless grad.
MB-STORM (Patel et al., 2022)	$\left(\frac{L\Delta\sigma}{S\sqrt{K}R}\right)^{2/3} + \left(\frac{L\Delta\zeta}{SR}\right)^{2/3} + \frac{\zeta^2}{SR} + \frac{L\Delta}{R} + \frac{\sigma^2}{NKR}$	Sample smooth Bounded hetero. $\mathcal{O}(K)$ batchsize[2]
CE-LSGD [1] (Patel et al., 2022)	$\left(\frac{L\Delta\sigma}{S\sqrt{K}R}\right)^{2/3} + \left(\frac{L\Delta\zeta}{SR}\right)^{2/3} + \frac{\zeta^2}{SR} + \frac{L\Delta}{R} + \frac{\sigma^2}{NKR}$	Sample smooth Bounded hetero. $\mathcal{O}(K)$ batchsize[2]
SCAFFOLD-M-VR (Thm. 4)	$\left(\frac{L\Delta\sigma}{S\sqrt{K}R}\right)^{2/3} + \frac{L\Delta}{R}\left(1+\frac{N^{1/2}}{S}\right)$	Sample smooth

[1] MIMELITEMVR and CE-LSGD consider the setting of streaming clients.

[2] A large batch is needed on each client per communication round.

1.2 Related work

Federated learning with homogeneous clients. FEDAVG is a well-known algorithm introduced by (McMahan et al., 2017b) as a heuristic to enhance communication efficiency and data privacy in federated learning. Numerous subsequent studies have focused on analyzing its convergence under the assumption of homogeneous datasets, where clients are independent and identically distributed (iid) and all clients participate fully (Stich, 2019; Yu et al., 2019b; Wang & Joshi, 2021; Lin et al., 2020; Zhou & Cong, 2017). However, when dealing with heterogeneous clients and partial client participation, FEDAVG is found to be vulnerable to data heterogeneity because of the ”client drift” effect (Karimireddy et al., 2020b; Yang et al., 2021; Wang et al., 2020b; Li et al., 2019).

Federated learning with heterogeneous clients. Considerable research efforts have been devoted to mitigating the impact of data heterogeneity in federated learning. For example, Li et al. (2020) propose FEDPROX, which introduces a proximal term to the objective function. Yang et al. (2021) utilize a two-sided learning rate approach, while Wang et al. (2020a) propose FEDNOVA, a normalized averaging method. Additionally, Zhang et al. (2021) present FEDPD, which addresses data heterogeneity from a primal-dual optimization perspective. Notably, Karimireddy et al. (2020b) introduces SCAFFOLD, an effective algorithm that employs control variables to mitigate the influence of data heterogeneity and partial client participation. FedGate (Haddadpour et al., 2021) and LED (Alghunaim, 2023) are two recent effective algorithms that have alleviated the impact of data heterogeneity, utilizing gradient tracking (Xu et al., 2015; Di Lorenzo & Scutari, 2016; Pu & Nedić, 2020; Xin et al., 2020; Alghunaim & Yuan, 2021) and exact-diffusion (Yuan et al., 2019, 2020, 2021a) techniques, respectively.

Federated learning with momentum. The momentum mechanism dates back to Nesterov's acceleration (Yurri, 2004) and Polyak's heavy-ball method (Polyak, 1964) in deterministic optimization, which later flourishes in the stochastic scenario (Yan et al., 2018; Yu et al., 2019a; Liu et al., 2020) and other communication efficient algorithms (Yuan et al., 2021b; He et al., 2023b,a). Extensive research has explored incorporating momentum into federated learning (Reddi et al., 2021; Wang et al., 2020b; Karimireddy et al., 2020a; Khanduri et al., 2021; Patel et al., 2022; Das et al., 2022; Yu et al., 2019a), and numerous empirical studies have demonstrated its substantial impact on enhancing the performance of federated learning algorithms (Wang et al., 2020b; Xu et al., 2021b; Reddi et al., 2021; Jin et al., 2022; Kim et al., 2022). However, whether momentum can offer *theoretical benefits* to federated learning remains underexplored. This work demonstrates that momentum can improve non-iid federated learning simply and provably. It is worth noting that the theoretical utility of momentum has been demonstrated in various scenarios beyond federated learning. For instance, Guo et al. (2021) proved that momentum can correct the bias experienced by the Adam method, while a very recent work (Fatkhullin et al., 2023) demonstrated that momentum can improve the error feedback technique in communication compression. The analysis presented in this work is different from (Guo et al., 2021) and (Fatkhullin et al., 2023) due to the unique challenges encountered in federated learning including multiple local updates, data heterogeneity, and partial client participation.

2 Problem setup

This section formulates the problem of non-iid federated learning. Formally, we consider minimizing the following objective with the fewest number of client-server communication rounds:

$$\min_{x\in\mathbb{R}^d} \quad f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x) \quad \text{where} \quad f_i(x) := \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[F(x;\xi_i)].$$

Here, the random variable ξ_i represents a local datapoint available at client i, while the function $f_i(x)$ denotes the non-convex local loss function associated with client i. This function takes expectation with respect to the local data distribution $\mathcal{D}_i$. In practice, the local data distributions $\mathcal{D}_i$ among different clients typically differ from each other, resulting in the inequality $f_i(x) \neq f_j(x)$ for any pair of nodes i and j. This phenomenon is commonly referred to as *data heterogeneity*. If all local clients were homogeneous, meaning that all local data samples follow the same distribution $\mathcal{D}$, we would have $f_i(x) = f_j(x)$ for any i and j. In addition, throughout the paper we assume that the function f is bounded from below and possesses a global minimum f^*. To facilitate convergence analysis, we also introduce the following standard assumptions.

Assumption 1 (Standard smoothness)**.** *Each local function $f_i(x)$ is differentiable. For any $1 \le i \le N$ and $x, \tilde{x} \in \mathbb{R}^d$, there exists a constant $L \ge 0$ such that*

$$\|\nabla f_i(x) - \nabla f_i(\tilde{x})\| \le L\|x - \tilde{x}\|.$$

Assumption 2 (Sample-wise smoothness)**.** *Each sample-wise function $F(x;\xi)$ is differentiable in terms of x. For any $1 \le i \le N$, $\xi_i \sim \mathcal{D}_i$, and $x, \tilde{x} \in \mathbb{R}^d$, there exists a constant $L \ge 0$ such that*

$$\|\nabla F(x;\xi_i) - \nabla F(\tilde{x};\xi_i)\| \le L\|x - \tilde{x}\|.$$

It is worth noting that Assumption 2 implies Assumption 1, which is typically used in variance-reduced algorithms, *e.g.*, (Karimireddy et al., 2020a; Khanduri et al., 2021; Fang et al., 2018; Cutkosky & Orabona, 2019). We will utilize either Assumption 1 or 2 in different algorithms.

Assumption 3 (Stochatic Gradient)**.** *For any $1 \le i \le N$ and $x \in \mathbb{R}^d$, there exists a constant $\sigma \ge 0$ such that*

$$\mathbb{E}_{\xi_i \sim \mathcal{D}_i}[\nabla F(x;\xi_i)] = \nabla f_i(x) \quad \textit{and} \quad \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[\|\nabla F(x,\xi_i) - \nabla f_i(x)\|^2] \le \sigma^2.$$

3 Accelerating FedAvg with momentum

This section focuses on the scenario in which all clients participate in the training process of federated learning. We will introduce momentum to both FedAvg and its variance-reduced extension. Additionally, we will demonstrate that the incorporation of momentum effectively eliminates the impact of data heterogeneity, leading to improved convergence rates.

3.1 FedAvg with momentum

Algorithm. We introduce momentum to enhance the estimation of the stochastic gradient, resulting in the new algorithm FedAvg-M, as presented in Algorithm 1. In FedAvg-M, the subscript i represents the client index, while the superscripts r and k denote the outer loop index and inner local update index, respectively. The structure of FedAvg-M remains identical to the vanilla FedAvg, except for the inclusion of momentum in gradient computation (see the highlight in Algorithm 1):

$$g_i^{r,k} = \beta \nabla F(x_i^{r,k};\xi_i^{r,k}) + (1-\beta)g^r,$$

where $\beta \in [0,1]$ is the momentum coefficient, and g^r represents an global gradient estimate updated in the outer loop r. It is important to note that FedAvg-M will reduce to the vanilla FedAvg when $\beta = 1$. Furthermore, FedAvg-M is easy to implement, as it maintains the same algorithmic structure and incurs no additional uplink communication overhead compared to FedAvg.

Algorithm 1 FedAvg-M: FedAvg with momentum

```
Require: initial model $x^0$ and gradient estimate $g^0$, local and global learning rate $\eta$ and $\gamma$, momentum $\beta$
  for $r = 0, \cdots, R-1$ do
    for each client $i \in \{1, \ldots, N\}$ in parallel do
      Initialize local model $x_i^{r,0} = x^r$
      for $k = 0, \cdots, K-1$ do
        Compute $g_i^{r,k} = \beta \nabla F(x_i^{r,k}; \xi_i^{r,k}) + (1-\beta) g^r$          ▷ $\beta = 1$ implies FedAvg
        Update local model $x_i^{r,k+1} = x_i^{r,k} - \eta g_i^{r,k}$
      end for
    end for
    Aggregate local updates $g^{r+1} = \frac{1}{\eta N K} \sum_{i=1}^{N} \left( x^r - x_i^{r,K} \right)$
    Update global model $x^{r+1} = x^r - \gamma g^{r+1}$
  end for
```

Convergence property. The inclusion of momentum in FedAvg yields notable theoretical improvements. Firstly, it eliminates the need for the data heterogeneity assumption, also known as the gradient similarity assumption. The assumption can be expressed as

$$\frac{1}{N} \sum_{i=1}^{N} \|\nabla f_i(x) - \nabla f(x)\|^2 \leq \zeta^2, \quad \forall x \in \mathbb{R}^d \qquad \text{(Data heterogeneity assumption)}$$

where ζ^2 measures the magnitude of data heterogeneity. By incorporating momentum, the above assumption is *no longer required* for the convergence analysis of FedAvg. Secondly, momentum enables FedAvg to converge at a state-of-the-art rate. These improvements are justified as follows:

Theorem 1. *Under Assumption 1 and 3, if we set $g^0 = 0$, $\beta = \min \left\{ 1, \sqrt{NKL\Delta/(\sigma^2 R)} \right\}$,*

$$\gamma = \min \left\{ \frac{1}{24L}, \frac{\beta}{6L} \right\}, \quad \eta K L \lesssim \min \left\{ 1, \frac{1}{\beta \gamma L R}, \left(\frac{L\Delta}{G_0 \beta^3 R} \right)^{1/2}, \frac{1}{(\beta N)^{1/2}}, \frac{1}{(\beta^3 N K)^{1/4}} \right\}$$

where $\Delta \triangleq f(x^0) - \min_x f(x)$ and $G_0 \triangleq N^{-1} \sum_{1 \leq i \leq N} \|\nabla f_i(x^0)\|^2$, then FedAvg-M *satisfies*

$$\frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E}[\|\nabla f(x^r)\|^2] \lesssim \sqrt{\frac{L\Delta\sigma^2}{NKR}} + \frac{L\Delta}{R}$$

where notation $\lesssim$ denotes inequalities that hold up to a numeric number.

Comparison with prior works. Table 1 compares FedAvg-M with existing algorithms when all clients participate in the training process. The results demonstrate that FedAvg-M attains the most favorable convergence rate without relying on any assumption of data heterogeneity. Moreover, this rate matches the lower bound provided by (Arjevani et al., 2019).

Constant local learning rate. Based on Theorem 1, it can be inferred that when $R \gtrsim NKL\Delta/\sigma^2$, FedAvg-M allows the utilization of *constant* local learning rate η which does not decay as the number of communication rounds R increases. This characteristic eases the tuning of the local learning rate and improves empirical performance. In contrast, many existing convergence results of FedAvg necessitate the adoption of local learning rates that diminish as R increases, as exemplified by *e.g.*, (Yang et al., 2021; Li et al., 2019; Karimireddy et al., 2020b; Koloskova et al., 2020).

Intuition on the effectiveness of momentum. The momentum mechanism relies on an accumulated gradient estimate g^r, which, although biased, exhibits reduced variance due to its accumulation nature compared to the stochastic gradient $\nabla F(x_i^{r,k};\xi_i^{r,k})$ computed with a single data batch. Importantly, by utilizing directions $\beta\nabla F(x_i^{r,k};\xi_i^{r,k})+(1-\beta)g^r$ for local updates, an "anchoring" effect is achieved, effectively mitigating the "client-drift" phenomenon. In the extreme case where $\beta = 0$, all clients remain synchronized in their local updates, eliminating any drift. By appropriately tuning the coefficient β, FedAvg-M maintains the same convergence rate as (Yang et al., 2021) while removing the requirement of data heterogeneity assumption utilized in their analysis.

3.2 Variance-reduced FedAvg with momentum

When each local loss function is further assumed to be sample-wise smooth (*i.e.*, Assumption 2), we can replace the local descent direction in Algorithm 1 with a variance-reduced momentum direction

$$g_i^{r,k} = \nabla F(x_i^{r,k};\xi_i^{r,k}) + (1-\beta)(g^r - \nabla F(x^{r-1};\xi_i^{r,k})) \tag{3.1}$$

to further enhance convergence, leading to variance-reduced FedAvg with momentum, or FedAvg-M-VR for short, see the detailed algorithm in Appendix B.2. The variable x^{r-1} is the last-iterate global model maintained in the server. The construction of the variance-reduced direction (3.1) effectively mitigates the influence of within-client gradient noise and can be traced back to SARAH (Nguyen et al., 2017) and STORM (Cutkosky & Orabona, 2019) in stochastic optimization; more discussion can be found in (Tan et al., 2022). Same as FedAvg-M, turning off the variance-reduced momentum of FedAvg-M-VR, *i.e.*, setting $\beta = 1$, recovers FedAvg. FedAvg-M-VR shares the same algorithmic structure and uplink communication workload as FedAvg.

Theorem 2. *Under Assumption 2 and 3, if we take* $g^0 = \frac{1}{NB}\sum_{i=1}^{N}\sum_{b=1}^{B}\nabla F(x^0;\xi_i^b)$ *with* $\{\xi_i^b\}_{b=1}^B \overset{iid}{\sim} \mathcal{D}_i$ *and set* $\beta = \min\left\{1, \left(\frac{NKL^2\Delta^2}{\sigma^4R^2}\right)^{1/3}\right\}$, $\gamma = \min\left\{\frac{1}{24L}, \sqrt{\frac{\beta NK}{54L^2}}\right\}$, $B = \left\lceil\frac{K}{R\beta^2}\right\rceil$, *and*

$$\eta KL \lesssim \min\left\{\left(\frac{L\Delta}{G_0\gamma LR}\right)^{1/2}, \left(\frac{\beta}{N}\right)^{1/2}, \left(\frac{\beta}{NK}\right)^{1/4}\right\},$$

then FedAvg-M-VR *converges as*

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}[\|\nabla f(x^r)\|^2] \lesssim \left(\frac{L\Delta\sigma}{NKR}\right)^{2/3} + \frac{L\Delta}{R}.$$

Comparison with prior works. FedAvg-M-VR surpasses all existing variance-reduced federated learning methods in terms of convergence rate, as demonstrated by the results presented in Table 1. Additionally, when compared to BVR-L-SGD (Murata & Suzuki, 2021) and CE-LSGD (Patel et al., 2022), FedAvg-M-VR computes each local stochastic gradient using a batchsize of $\mathcal{O}(1)$, contrasting with the $O(K)$ batchsize employed by BVR-L-SGD and CE-LSGD. Furthermore, in comparison to STEM (Khanduri et al., 2021), FedAvg-M-VR does not rely on the assumption of bounded data heterogeneity.

Based on discussions in Sections 3.1 and 3.2, we demonstrate that FedAvg-M and FedAvg-M-VR, in the context of full client participation, can achieve the state-of-the-art convergence rate without resorting to any stronger assumption, *e.g.*, bounded data heterogeneity or impractical algorithmic structures such as large batchsizes.

4 Accelerating SCAFFOLD with momentum

This section addresses the scenario where a random subset of clients participates in the training process per iteration. To tackle the challenges arising from partial participation, SCAFFOLD employs a control variable in each client to counteract the "client drift" effect during local updates. To further enhance the convergence performance, we will introduce momentum to both SCAFFOLD and its variance-reduced extension. Through our analysis, we will demonstrate that the incorporation of momentum results in new state-of-the-art convergence rates for these algorithms.

4.1 SCAFFOLD with momentum

Algorithm. We introduce momentum to enhance the estimation of the stochastic gradient, resulting in the newly proposed algorithm SCAFFOLD-M, outlined in Algorithm 2. In SCAFFOLD-M, S clients are randomly selected from a pool of N clients for each iteration of trainining. The control variables c_i and c are maintained by the client and server, respectively. In SCAFFOLD, the local descent direction is given by $\nabla F(x_i^{r,k};\xi_i^{r.k}) - c_i^r + c^r$. In contrast, SCAFFOLD-M incorporates momentum directions for local updates:

$$g_i^{r,k} = \beta(\nabla F(x_i^{r,k};\xi_i^{r,k}) - c_i^r + c^r) + (1-\beta)g^r,$$

where g^r represents the global stochastic gradient vector maintained by the server. It is worth noting that SCAFFOLD-M can reduce to SCAFFOLD by setting $\beta = 1$.

Algorithm 2 SCAFFOLD-M: SCAFFOLD with momentum

```
Require: initial model $x^0$, gradient estimator $g^0$, control variables $\{c_i^0\}_{i=1}^N$ and $c^0$, local learning rate $\eta$,
  global learning rate $\gamma$, momentum $\beta$
  for $r = 0, \cdots, R-1$ do
    Uniformly sample clients $\mathcal{S}_r \subseteq \{1, \cdots, N\}$ with $|\mathcal{S}_r| = S$
    for each client $i \in \mathcal{S}_r$ in parallel do
      Initialize local model $x_i^{r,0} = x^r$
      for $k = 0, \cdots, K-1$ do
        Compute $g_i^{r,k} = \beta(\nabla F(x_i^{r,k}; \xi_i^{r,k}) - c_i^r + c^r) + (1-\beta) g^r$        ▷ $\beta = 1$ implies SCAFFOLD
        Update local model $x_i^{r,k+1} = x_i^{r,k} - \eta g_i^{r,k}$
      end for
      Update control variable $c_i^{r+1} := \frac{1}{K}\sum_{k=0}^{K-1} \nabla F(x_i^{r,k}; \xi_i^{r,k})$ (for $i \notin \mathcal{S}_r$, $c_i^{r+1} = c_i^r$)
    end for
    Aggregate local updates $g^{r+1} = \frac{1}{\eta SK}\sum_{i\in\mathcal{S}_r}\left(x^r - x_i^{r,K}\right)$
    Update global model $x^{r+1} = x^r - \gamma g^{r+1}$
    Update control variable $c^{r+1} = c^r + \frac{1}{N}\sum_{i\in\mathcal{S}_r}(c_i^{r+1} - c_i^r)$
  end for
```

Convergence property. The inclusion of momentum in SCAFFOLD yields notable theoretical improvements, as justified by the following theorem.

Theorem 3. *Under Assumption 1 and 3, if we take* $g^0 = 0$, $c_i^0 = \frac{1}{B}\sum_{b=1}^{B} \nabla F(x^0; \xi_i^b)$ *with* $\{\xi_i^b\}_{b=1}^B \overset{iid}{\sim} \mathcal{D}_i$, $c^0 = \frac{1}{N}\sum_{i=1}^{N} c_i^0$ *and set* $B = \left\lceil \frac{NK}{SR} \right\rceil$, $\gamma = \frac{\beta}{L}$, $\beta = \min\left\{1, \frac{S}{N^{2/3}}, \sqrt{\frac{L\Delta SK}{\sigma^2 R}}, \sqrt{\frac{L\Delta S^2}{G_0 N}}\right\}$, $\eta KL \lesssim \min\left\{\frac{1}{S^{1/2}}, \frac{1}{\beta K^{1/4}}, \frac{S^{1/2}}{N}\right\}$, *then* SCAFFOLD-M *converges as*

$$\frac{1}{R}\sum_{r=0}^{R-1} \mathbb{E}[\|\nabla f(x^r)\|^2] \lesssim \sqrt{\frac{L\Delta\sigma^2}{SKR}} + \frac{L\Delta}{R}\left(1 + \frac{N^{2/3}}{S}\right).$$

Comparison with SCAFFOLD**.** Compared to SCAFFOLD, SCAFFOLD-M exhibits provably faster convergence under partial participation, as demonstrated in the comparison presented in Table 2. Specifically, when the gradients are noiseless (*i.e.*, $\sigma^2 = 0$), achieving the same level of precision requires a ratio, between SCAFFOLD-M and SCAFFOLD, of communication rounds given by

$$\frac{1 + N^{2/3}/S}{(N/S)^{2/3}} \asymp \max\left\{\left(\frac{S}{N}\right)^{2/3}, \frac{1}{S^{1/3}}\right\},$$

where notation $\asymp$ indicates that the equality holds up to a numeric number. Consequently, if $S \asymp N^{2/3}$, SCAFFOLD-M achieves up to $N^{2/9}$ times improvement in comparison to the vanilla SCAFFOLD, when aiming for the same precision. This improvement is particularly significant as N, the number of clients, is typically very large. It is also worth noting that prior to the introduction of SCAFFOLD-M, SCAFFOLD was the only known non-iid federated learning method, to the best of our knowledge, that is robust to both unbounded data heterogeneity and partial client sampling, and capable of attaining linear speedup without relying on impractical algorithmic structures. Consequently, the development of SCAFFOLD-M provides an alternative and superior choice.

4.2 Variance-reduced SCAFFOLD with momentum

Similar to FedAvg-M-VR, when the loss functions further enjoy the sample-wise smoothness property, we can obtain SCAFFOLD-M-VR by replacing momentum directions in Algorithm 2 with variance-reduced momentum directions

$$g_i^{r,k} = \nabla F(x_i^{r,k};\xi_i^{r,k}) - \beta(c_i^r - c^r) + (1-\beta)(g^r - \nabla F(x^{r-1};\xi_i^{r,k})).$$

The detailed algorithm is in Appendix C.2, and the convergence is shown below.

Theorem 4. *Under Assumption 2 and 3, if we take* $c_i^0 = \frac{1}{B}\sum_{b=1}^{B}\nabla F(x^0;\xi_i^b)$ *with* $\{\xi_i^b\}_{b=1}^{B} \overset{iid}{\sim} \mathcal{D}_i$, $g^0 = c^0 = \frac{1}{N}\sum_{i=1}^{N} c_i^0$ *and set* $\beta = \min\left\{\frac{S}{N}, \left(\frac{KL\Delta}{\sigma^2 R}\right)^{2/3} S^{1/3}\right\}$, $\gamma = \min\left\{\frac{1}{L}, \frac{\sqrt{\beta S}}{L}\right\}$, $B = \left\lceil \max\left\{\frac{SK}{NR\beta^2}, \frac{NK}{SR}\right\}\right\rceil$, $\eta KL \lesssim \min\left\{\left(\frac{\beta}{S}\right)^{1/2}, \left(\frac{\beta}{SK}\right)^{1/4}\right\}$, *then* SCAFFOLD-M-VR *converges as*

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}[\|\nabla f(x^r)\|^2] \lesssim \left(\frac{L\Delta\sigma}{S\sqrt{K}R}\right)^{2/3} + \frac{L\Delta}{R}\left(1 + \frac{N^{1/2}}{S}\right).$$

Comparison with variance-reduced methods. SCAFFOLD-M-VR outperforms all existing variance-reduced federated learning methods under partial participation in terms of convergence rate when data heterogeneity is severe (*i.e.*, ζ^2 is large), see results listed in Table 2. Moreover, SCAFFOLD-M-VR has the following additional advantages. Compared to MimeLiteMVR (Karimireddy et al., 2020a), SCAFFOLD-M-VR does not need access to noiseless (full-batch) local gradients per iteration. Compared to MB-STORM (Patel et al., 2022) and CE-LSGD (Patel et al., 2022), SCAFFOLD-M-VR does not require bounded data heterogeneity and computes each gradient efficiently with batchsize 1, as opposed to batchsize $\mathcal{O}(K)$.

Based on discussions in Sections 4.1 and 4.2, we demonstrate that SCAFFOLD-M and SCAFFOLD-M-VR, in the context of partial client participation, can achieve state-of-the-art convergence rates without resorting to any stronger assumption, *e.g.*, bounded data heterogeneity or impractical algorithmic structures such as large batchsize.

5 Experiments

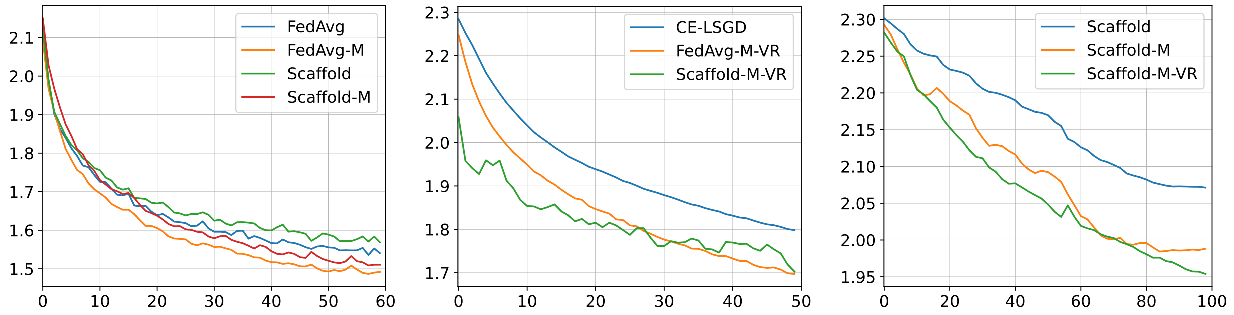
Experimental settings. We conducted an evaluation of our proposed methods using a three-layer fully connected neural network trained on the CIFAR-10 dataset. To generate non-iid data for the clients, we sample label ratios from the Dirichlet distribution (Hsu et al., 2019) with a parameter of 0.5 for the full participation setting and 0.2 for the partial participation setting. Our experimental setup involves $N = 10$ clients and $K = 32$ local updates. The weight decay is set as 10^{-4}. The global learning rate is fixed as $\gamma = \eta K$ for all the algorithms, and we performe a grid search for the local learning rate η in values $\{0.005, 0.01, 0.05, 0.1, 0.5\}$. Similarly, we search for the momentum parameter β in values $\{0.1, 0.2, 0.5, 0.8\}$.

Experimental results. Our experiments can be categorized into three parts.

Firstly, we compare the performance of FedAvg-M and SCAFFOLD-M with their momentumless counterparts, namely the vanilla FedAvg and SCAFFOLD, under full client participation. The results are presented in Figure 1(a), where it can be observed that incorporating momentum significantly accelerates the convergence of both FedAvg and SCAFFOLD.

Secondly, we compare three momentum-based variance-reduced methods: CE-LSGD, FedAvg-M-VR, and SCAFFOLD-M-VR, under the condition of full client participation. The comparison is illustrated in Figure 1(b). It is evident that our proposed methods outperform CE-LSGD with substantial margins.

Lastly, we investigate the partial participation setting with $S = 1$ and compare the performance of SCAFFOLD-M and SCAFFOLD-M-VR with vanilla SCAFFOLD. The results are presented in Figure 1(c). Once again, we observe that the introduction of momentum leads to significant improvements even when only a few clients participate in each round of training.



(a) Acceleration of momentum

(b) Comparison of VR methods

(c) Partial participation

Figure 1: Test loss versus the number of communication rounds

6 Conclusion

This paper proposes momentum variants of FedAvg and SCAFFOLD under various client participation situations and smoothness properties. All of our momentum variants only make simple and practical modifications to FedAvg and SCAFFOLD yet obtain state-of-the-art performance among their peers, particularly when data heterogeneity is severe or gradient noise is trivial. In particular, FedAvg-M converges without relying on bounded data heterogeneity and can adopt constant local learning rates, giving the *first* neat convergence guarantee for FedAvg-type methods; SCAFFOLD-M is the *first* FL method that outperforms SCAFFOLD unconditionally. Experiments are conducted in the paper to support our theoretical findings.

References

Sulaiman A Alghunaim. Local exact-diffusion for decentralized optimization and learning. *arXiv:2302.00620*, 2023.

Sulaiman A Alghunaim and Kun Yuan. A unified and refined convergence analysis for non-convex decentralized learning. *arXiv preprint arXiv:2110.09993*, 2021.

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

Tianyi Chen, Xiao Jin, Yuejiao Sun, and Wotao Yin. Vafl: a method of vertical asynchronous federated learning. *arXiv preprint arXiv:2007.06081*, 2020a.

Yujing Chen, Yue Ning, Martin Slawski, and Huzefa Rangwala. Asynchronous online federated learning for edge devices with non-iid data. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 15–24. IEEE, 2020b.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

Rudrajit Das, Anish Acharya, Abolfazl Hashemi, Sujay Sanghavi, Inderjit S Dhillon, and Ufuk Topcu. Faster non-convex federated learning via global and local momentum. In *Uncertainty in Artificial Intelligence*, pp. 496–506. PMLR, 2022.

P. Di Lorenzo and G. Scutari. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.

Alp Emre Durmus, Zhao Yue, Matas Ramon, Mattina Matthew, Whatmough Paul, and Saligrama Venkatesh. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. In *Advances in Neural Information Processing Systems*, 2018.

Ilyas Fatkhullin, Alexander Tyurin, and Peter Richtárik. Momentum provably improves error feedback! *arXiv preprint arXiv:2305.15155*, 2023.

Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. A novel convergence analysis for algorithms of the adam family. *arXiv preprint arXiv:2112.03459*, 2021.

Farzin Haddadpour, Mohammad Mahdi Kamani, Aryan Mokhtari, and Mehrdad Mahdavi. Federated learning with compression: Unified analysis and sharp guarantees. In *International Conference on Artificial Intelligence and Statistics*, pp. 2350–2358. PMLR, 2021.

Yutong He, Xinmeng Huang, Yiming Chen, Wotao Yin, and Kun Yuan. Lower bounds and accelerated algorithms in distributed stochastic optimization with communication compression. *arXiv preprint arXiv:2305.07612*, 2023a.

Yutong He, Xinmeng Huang, and Kun Yuan. Unbiased compression saves communication in distributed optimization: When and how much? *arXiv preprint arXiv:2305.16297*, 2023b.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Jiayin Jin, Jiaxiang Ren, Yang Zhou, Lingjuan Lyu, Ji Liu, and Dejing Dou. Accelerated federated learning with decoupled adaptive optimization. In *International Conference on Machine Learning*, pp. 10298–10322. PMLR, 2022.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv:2008.03606*, 2020a.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020b.

Prashant Khanduri, Pranay Sharma, Haibo Yang, Mingyi Hong, Jia Liu, Ketan Rajawat, and Pramod Varshney. Stem: A stochastic two-sided momentum algorithm achieving near-optimal sample and communication complexities for federated learning. *Advances in Neural Information Processing Systems*, 34: 6050–6061, 2021.

Geeho Kim, Jinkyu Kim, and Bohyung Han. Communication-efficient federated learning with acceleration of global momentum. *arXiv:2201.03172*, 2022.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pp. 5381–5393. PMLR, 2020.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2019.

Xianfeng Liang, Shuheng Shen, Jingchang Liu, Zhen Pan, Enhong Chen, and Yifei Cheng. Variance reduced local sgd with lower communication complexity. *arXiv preprint arXiv:1912.12844*, 2019.

Tao Lin, Sebastian U Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't use large mini-batches, use local sgd. In *International Conference on Learning Representations*, 2020.

Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. *Advances in Neural Information Processing Systems*, 33:18261–18271, 2020.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, 2017a.

H. B. McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2017b.

Aritra Mitra, Rayana Jaafar, George J Pappas, and Hamed Hassani. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. *Advances in Neural Information Processing Systems*, 34:14606–14619, 2021.

Tomoya Murata and Taiji Suzuki. Bias-variance reduced local sgd for less heterogeneous federated learning. In *International Conference on Machine Learning*, pp. 7872–7881. PMLR, 2021.

Edward Duc Hien Nguyen, Sulaiman A Alghunaim, Kun Yuan, and César A Uribe. On the performance of gradient tracking with local updates. *arXiv:2210.04757*, 2022.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pp. 2613–2621. PMLR, 2017.

Kumar Kshitij Patel, Lingxiao Wang, Blake E Woodworth, Brian Bullins, and Nati Srebro. Towards optimal communication complexity in distributed non-convex optimization. *Advances in Neural Information Processing Systems*, 35:13316–13328, 2022.

Krishna Pillutla, Kshitiz Malik, Abdel-Rahman Mohamed, Mike Rabbat, Maziar Sanjabi, and Lin Xiao. Federated learning with partial model personalization. In *International Conference on Machine Learning*, pp. 17716–17758. PMLR, 2022.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

Shi Pu and Angelia Nedić. Distributed stochastic gradient tracking methods. *Mathematical Programming*, pp. 1–49, 2020.

Sashank J Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2021–2031. PMLR, 2020.

Elsa Rizk, Stefan Vlaski, and Ali H Sayed. Privatized graph federated learning. *arXiv:2203.07105*, 2022.

Sebastian Urban Stich. Local sgd converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of local-update sgd algorithms. *The Journal of Machine Learning Research*, 22(1):9709–9758, 2021.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33: 7611–7623, 2020a.

Jianyu Wang, Vinayak Tantia, Nicolas Ballas, and Michael Rabbat. SlowMo: Improving communication-efficient distributed sgd with slow momentum. In *International Conference on Learning Representations*, 2020b.

Ran Xin, Usman A Khan, and Soummya Kar. An improved convergence analysis for decentralized online stochastic non-convex optimization. *IEEE Transactions on Signal Processing*, 2020.

Chenhao Xu, Youyang Qu, Yong Xiang, and Longxiang Gao. Asynchronous federated learning on heterogeneous devices: A survey. *arXiv preprint arXiv:2109.04269*, 2021a.

J. Xu, S. Zhu, Y. C. Soh, and L. Xie. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *IEEE Conference on Decision and Control (CDC)*, pp. 2055–2060, Osaka, Japan, 2015.

Jing Xu, Sen Wang, Liwei Wang, and Andrew Chi-Chih Yao. FedCM: Federated learning with client-level momentum. *arXiv:2106.10874*, 2021b.

Yan Yan, Tianbao Yang, Zhe Li, Qihang Lin, and Yi Yang. A unified analysis of stochastic momentum methods for deep learning. In *International Joint Conference on Artificial Intelligence*, 2018.

Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-iid federated learning. In *International Conference on Learning Representations*, 2021.

Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. In *International Conference on Machine Learning*, pp. 7184–7193. PMLR, 2019a.

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

Kun Yuan, Bicheng Ying, Xiaochuan Zhao, and Ali H. Sayed. Exact diffusion for distributed optimization and learning—part i: Algorithm development. *IEEE Transactions on Signal Processing*, 67:708–723, 2019.

Kun Yuan, Sulaiman A Alghunaim, Bicheng Ying, and Ali H Sayed. On the influence of bias-correction on distributed stochastic optimization. *IEEE Transactions on Signal Processing*, 2020.

Kun Yuan, Sulaiman A. Alghunaim, and Xinmeng Huang. Removing data heterogeneity influence enhances network topology dependence of decentralized sgd. *ArXiv*, abs/2105.08023, 2021a.

Kun Yuan, Yiming Chen, Xinmeng Huang, Yingya Zhang, Pan Pan, Yinghui Xu, and Wotao Yin. DecentLaM: Decentralized momentum sgd for large-batch deep training. *International Conference on Computer Vision*, pp. 3009–3019, 2021b.

Nesterov Yurri. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, Norwell, 2004.

Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu. FedPD: A federated learning framework with adaptivity to non-iid data. *IEEE Transactions on Signal Processing*, 69:6055–6070, 2021.

Fan Zhou and Guojing Cong. On the convergence properties of a k-step averaging stochastic gradient descent algorithm for nonconvex optimization. *arXiv preprint arXiv:1708.01012*, 2017.

A Preliminaries of proofs

Let $\mathcal{F}^0 = \emptyset$ and $\mathcal{F}_i^{r,k} := \sigma(\{x_i^{r,j}\}_{0\le j\le k} \cup \mathcal{F}^r)$ and $\mathcal{F}^{r+1} := \sigma(\cup_i \mathcal{F}_i^{r,K})$ for all $r \ge 0$ where $\sigma(\cdot)$ indicates the σ-algebra. Let $\mathbb{E}_r[\cdot] := \mathbb{E}[\cdot|\mathcal{F}^r]$ be the expectation, conditioned on the filtration $\mathcal{F}^r$, with respect to the random variables $\{\mathcal{S}^r, \{\xi_i^{r,k}\}_{1\le i\le N, 0\le k<K}\}$ in the r-th iteration. We also use $\mathbb{E}[\cdot]$ to denote the global expectation over all randomness in algorithms. Through out the proofs, we use $\sum_i$ to represent the sum over $i \in \{1, \dots, N\}$, while $\sum_{i\in\mathcal{S}^r}$ denotes the sum over $i \in \mathcal{S}^r$. Similarly, we use $\sum_k$ to represent the sum over $k \in \{0, \dots, K-1\}$. For all $r \ge 0$, we define the following auxiliary variables to facilitate proofs:

$$\begin{aligned}
\mathcal{E}_r &:= \mathbb{E}[\|\nabla f(x^r) - g^{r+1}\|^2], \\
U_r &:= \frac{1}{NK}\sum_i\sum_k \mathbb{E}[\|x_i^{r,k} - x^r\|]^2, \\
\zeta_i^{r,k} &:= \mathbb{E}[x_i^{r,k+1} - x_i^{r,k}|\mathcal{F}_i^{r,k}], \\
\Xi_r &:= \frac{1}{N}\sum_{i=1}^N \mathbb{E}[\|\zeta_i^{r,0}\|^2], \\
V_r &:= \frac{1}{N}\sum_{i=1}^N \mathbb{E}[\|c_i^r - \nabla f_i(x^{r-1})\|^2].
\end{aligned}$$

We remark that quantity V_r is only used in the analysis of SCAFFOLD-based algorithms. Throughout the appendix, we let $\Delta := f(x^0) - f^*, G_0 := \frac{1}{N}\sum_i \|\nabla f_i(x^0)\|^2$, $x^{-1} := x^0$ and $\mathcal{E}_{-1} := \mathbb{E}[\|\nabla f(x^0) - g^0\|^2]$. We will use the following fundamental lemmas for all our algorithms.

Lemma A.1. *Under Assumption 1, if $\gamma L \le \frac{1}{24}$, the following inequality holds for all $r \ge 0$:*

$$\mathbb{E}[f(x^{r+1})] \le \mathbb{E}[f(x^r)] - \frac{11\gamma}{24}\mathbb{E}[\|\nabla f(x^r)\|^2] + \frac{13\gamma}{24}\mathcal{E}_r.$$

Proof. Since f is L-smooth, we have

$$\begin{aligned}
f(x^{r+1})] \le& f(x^r) + \langle \nabla f(x^r), x^{r+1} - x^r\rangle + \frac{L}{2}\|x^{r+1} - x^r\|^2 \\
=& f(x^r) - \gamma\|\nabla f(x^r)\|^2 + \gamma\langle \nabla f(x^r), g^{r+1}\rangle + \frac{L\gamma^2}{2}\|g^{r+1}\|^2.
\end{aligned}$$

Since $x^{r+1} = x^r - \gamma g^{r+1}$, using Young's inequality, we further have

$$\begin{aligned}
& f(x^{r+1}) \\
\le& f(x^r) - \frac{\gamma}{2}\|\nabla f(x^r)\|^2 + \frac{\gamma}{2}\|\nabla f(x^r) - g^{r+1}\|^2 + L\gamma^2(\|\nabla f(x^r)\|^2 + \|\nabla f(x^r) - g^{r+1}\|^2) \\
\le& f(x^r) - \frac{11\gamma}{24}\|\nabla f(x^r)\|^2 + \frac{13\gamma}{24}\|\nabla f(x^r) - g^{r+1}\|^2,
\end{aligned}$$

where the last inequality holds due to $\gamma L \le \frac{1}{24}$. Taking the global expectation completes the proof. □

To handle local updates and client sampling, we will also use the following technical lemmas.

Lemma A.2 (AM-GM inequality)**.** *Let $\{v_1, \cdots, v_\tau\}$ be τ vectors in $\mathbb{R}^d$. Then the following are true:*

1. $\|v_i + v_j\|^2 \le (1+a)\|v_i\|^2 + \left(1 + \frac{1}{a}\right)\|v_j\|^2$ *for any $a > 0$,*

2. $\left\|\sum_{i=1}^{\tau} v_i\right\|^2 \le \tau \sum_{i=1}^{\tau} \|v_i\|^2$.

Lemma A.3 (Karimireddy et al. (2020b))**.** *Suppose $\{X_1, \cdots, X_\tau\} \subset \mathbb{R}^d$ be random variables that are potentially dependent. If their marginal means and variances satisfy $\mathbb{E}[X_i] = \mu_i$ and $\mathbb{E}[\|X_i - \mu_i\|^2] \le \sigma^2$, then it holds that*

$$\mathbb{E}\left[\left\|\sum_{i=1}^{\tau} X_i\right\|^2\right] \le \left\|\sum_{i=1}^{\tau} \mu_i\right\|^2 + \tau^2\sigma^2.$$

If they are correlated in the Markov way such that $\mathbb{E}[X_i | X_{i-1}, \cdots X_1] = \mu_i$ and $\mathbb{E}[\|X_i - \mu_i\|^2] \le \sigma^2$, i.e., the variables $\{X_i - \mu_i\}$ form a martingale. Then the following tighter bound holds:

$$\mathbb{E}\left[\left\|\sum_{i=1}^{\tau} X_i\right\|^2\right] \le 2\mathbb{E}\left[\left\|\sum_{i=1}^{\tau} \mu_i\right\|^2\right] + 2\tau\sigma^2.$$

Lemma A.4. *Given vectors $v_1, \cdots, v_N \in \mathbb{R}^d$ and $\bar{v} = \frac{1}{N}\sum_{i=1}^{N} v_i$, if we sample $\mathcal{S} \subset \{1, \cdots, N\}$ uniformly randomly such that $|\mathcal{S}| = S$, then it holds that*

$$\mathbb{E}\left[\left\|\frac{1}{S}\sum_{i\in\mathcal{S}} v_i\right\|^2\right] = \|\bar{v}\|^2 + \frac{N-S}{S(N-1)}\frac{1}{N}\sum_{i=1}^{N}\|v_i - \bar{v}\|^2.$$

Proof. Letting $\mathbb{1}\{i \in \mathcal{S}\}$ be the indicator for the event $i \in \mathcal{S}_r$, we prove this lemma as follows:

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{S}\sum_{i\in\mathcal{S}} v_i\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{S}\sum_{i=1}^{N} v_i \mathbb{1}\{i \in \mathcal{S}\}\right\|^2\right] \\
&= \frac{1}{S^2}\mathbb{E}\left[\left(\sum_{i} \|v_i\|^2 \mathbb{1}\{i \in \mathcal{S}\} + 2\sum_{i<j} v_i^\top v_j \mathbb{1}\{i, j \in \mathcal{S}\}\right)\right] \\
&= \frac{1}{SN}\sum_{i=1}^{N}\|v_i\|^2 + \frac{1}{S^2}\frac{S(S-1)}{N(N-1)}2\sum_{i<j} v_i^\top v_j \\
&= \frac{1}{SN}\sum_{i=1}^{N}\|v_i\|^2 + \frac{1}{S^2}\frac{S(S-1)}{N(N-1)}\left(\left\|\sum_{i=1}^{N} v_i\right\|^2 - \sum_{i=1}^{N}\|v_i\|^2\right) \\
&= \frac{N-S}{S(N-1)}\frac{1}{N}\sum_{i=1}^{N}\|v_i\|^2 + \frac{N(S-1)}{S(N-1)}\|\bar{v}\|^2 \\
&= \frac{N-S}{S(N-1)}\frac{1}{N}\sum_{i=1}^{N}\|v_i - \bar{v}\|^2 + \|\bar{v}\|^2.
\end{aligned}$$

□

In the following subsections, we present complete proofs of our main results. For FedAvg-M and SCAFFOLD-M, our proofs only rely on Assumption 1 and 3, while for FedAvg-M-VR and SCAFFOLD-M-VR, our proofs rely on Assumption 2 and 3.

B FedAvg with momentm

B.1 FedAvg-M

Lemma B.1. *If $\gamma L \le \frac{\beta}{6}$, the following holds for $r \ge 1$:*

$$\mathcal{E}_r \le \left(1 - \frac{8\beta}{9}\right)\mathcal{E}_{r-1} + \frac{4\gamma^2 L^2}{\beta}\mathbb{E}[\|\nabla f(x^{r-1})\|^2] + \frac{2\beta^2\sigma^2}{NK} + 4\beta L^2 U_r.$$

Additionally, it holds for $r = 0$ that

$$\mathcal{E}_0 \le (1-\beta)\mathcal{E}_{-1} + \frac{2\beta^2\sigma^2}{NK} + 4\beta L^2 U_0.$$

Proof. For $r \ge 1$,

$$\begin{aligned}
\mathcal{E}_r &= \mathbb{E}[\|\nabla f(x^r) - g^{r+1}\|^2] \\
&= \mathbb{E}\left[\left\|(1-\beta)(\nabla f(x^r) - g^r) + \beta\left(\nabla f(x^r) - \frac{1}{NK}\sum_i\sum_k \nabla F(x_i^{r,k};\xi_i^{r,k})\right)\right\|^2\right] \\
&= \mathbb{E}\left[\|(1-\beta)(\nabla f(x^r) - g^r)\|^2\right] + \beta^2\mathbb{E}\left[\left\|\nabla f(x^r) - \frac{1}{NK}\sum_{i,k}\nabla F(x_i^{r,k};\xi_i^{r,k})\right\|^2\right] \\
&\quad + 2\beta\mathbb{E}\left[\left\langle (1-\beta)(\nabla f(x^r) - g^r), \nabla f(x^r) - \frac{1}{NK}\sum_{i,k}\nabla f(x_i^{r,k})\right\rangle\right].
\end{aligned}$$

Note that $\{\nabla F(x_i^{r,k};\xi_i^{r,k})\}_{0\le k<K}$ are sequentially correlated. Applying the AM-GM inequality and Lemma A.3, we have

$$\mathcal{E}_r \le \left(1 + \frac{\beta}{2}\right)\mathbb{E}[\|(1-\beta)(\nabla f(x^r) - g^r)\|^2] + 2\beta L^2 U_r + 2\beta^2\left(\frac{\sigma^2}{NK} + L^2 U_r\right).$$

Using the AM-GM inequality again and Assumption 1, we have

$$\begin{aligned}
\mathcal{E}_r &\le (1-\beta)^2\left(1+\frac{\beta}{2}\right)\left[\left(1+\frac{\beta}{2}\right)\mathcal{E}_{r-1} + \left(1 + \frac{2}{\beta}\right)L^2\mathbb{E}[\|x^r - x^{r-1}\|^2]\right] + \frac{2\beta^2\sigma^2}{NK} + 4\beta L^2 U_r \\
&\le (1-\beta)\mathcal{E}_{r-1} + \frac{2}{\beta}L^2\mathbb{E}[\|x^r - x^{r-1}\|^2] + \frac{2\beta^2\sigma^2}{NK} + 4\beta L^2 U_r \\
&\le \left(1 - \frac{8\beta}{9}\right)\mathcal{E}_{r-1} + 4\frac{\gamma^2 L^2}{\beta}\mathbb{E}[\|\nabla f(x^{r-1})\|^2] + \frac{2\beta^2\sigma^2}{NK} + 4\beta L^2 U_r,
\end{aligned}$$

where we plug in $\|x^r - x^{r-1}\|^2 \le 2\gamma^2(\|\nabla f(x^{r-1})\|^2 + \|g^r - \nabla f(x^{r-1})\|^2)$ and use $\gamma L \le \frac{\beta}{6}$ in the last inequality. Similarly, for $r = 0$,

$$\begin{aligned}
\mathcal{E}_0 &\le \left(1 + \frac{\beta}{2}\right)\mathbb{E}[\|(1-\beta)(\nabla f(x^0) - g^0)\|^2] + 2\beta L^2 U_0 + 2\beta^2\left(\frac{\sigma^2}{NK} + L^2 U_0\right) \\
&\le (1-\beta)\mathcal{E}_{-1} + \frac{2\beta^2\sigma^2}{NK} + 4\beta L^2 U_0.
\end{aligned}$$

□

Lemma B.2. *If* $\eta LK \leq \dfrac{1}{\beta}$*, the following holds for* $r \geq 0$*:*

$$U_r \leq 2eK^2\Xi_r + K\eta^2\beta^2\sigma^2(1 + 2K^3L^2\eta^2\beta^2).$$

Proof. Recall that $\zeta_i^{r,k} := \mathbb{E}[x_i^{r,k+1} - x_i^{r,k} | \mathcal{F}_i^{r,k}] = -\eta\left((1-\beta)g^r + \beta\nabla f_i(x_i^{r,k})\right)$. Then we have

$$\begin{aligned}\mathbb{E}[\|\zeta_i^{r,j} - \zeta_i^{r,j-1}\|^2] &\leq \eta^2L^2\beta^2\mathbb{E}[\|x_i^{r,j} - x_i^{r,j-1}\|^2] \\ &\leq \eta^2L^2\beta^2(\eta^2\beta^2\sigma^2 + \mathbb{E}[\|\zeta_i^{r,j-1}\|^2).\end{aligned}$$

For any $1 \leq j \leq k-1 \leq K-2$, using $\eta L \leq \dfrac{1}{\beta K} \leq \dfrac{1}{\beta(k+1)}$, we have

$$\begin{aligned}\mathbb{E}[\|\zeta_i^{r,j}\|^2] &\leq \left(1 + \frac{1}{k}\right)\mathbb{E}[\|\zeta_i^{r,j-1}\|^2] + (1+k)\mathbb{E}[\|\zeta_i^{r,j} - \zeta_i^{r,j-1}\|^2] \\ &\leq \left(1 + \frac{2}{k}\right)\mathbb{E}[\|\zeta_i^{r,j-1}\|^2] + (k+1)L^2\eta^4\beta^4\sigma^2 \\ &\leq e^2\mathbb{E}[\|\zeta_i^{r,0}\|^2] + 4k^2L^2\eta^4\beta^4\sigma^2,\end{aligned}$$

where the last inequality holds by unrolling the recursive bound and using $\left(1 + \dfrac{2}{k}\right)^k \leq e^2$. By Lemma A.3, it holds that for $k \geq 2$,

$$\begin{aligned}\mathbb{E}[\|x_i^{r,k} - x^r\|^2] &\leq 2\mathbb{E}\left[\left\|\sum_{j=0}^{k-1}\zeta_i^{r,j}\right\|^2\right] + 2k\eta^2\beta^2\sigma^2 \\ &\leq 2k\sum_{j=0}^{k-1}\mathbb{E}[\|\zeta_i^{r,k}\|^2] + 2k\eta^2\beta^2\sigma^2 \\ &\leq 2e^2k^2\mathbb{E}[\|\zeta_i^{r,0}\|^2] + 2k\eta^2\beta^2\sigma^2(1 + 4k^3L^2\eta^2\beta^2).\end{aligned}$$

This is also valid for $k = 0, 1$. Summing up over i and k finishes the proof. □

Lemma B.3. *If* $288e(\eta KL)^2((1-\beta)^2 + e(\beta\gamma LR)^2) \leq 1$*, then it holds for* $r \geq 0$ *that*

$$\sum_{r=0}^{R-1}\Xi_r \leq \frac{1}{72eK^2L^2}\sum_{r=-1}^{R-2}(\mathcal{E}_r + \mathbb{E}[\|\nabla f(x^r)\|^2]) + 2\eta^2\beta^2eRG_0.$$

Proof. Note that $\zeta_i^{r,0} = -\eta((1-\beta)g^r + \beta\nabla f_i(x^r))$,

$$\frac{1}{N}\sum_{i=1}^{N}\|\zeta_i^{r,0}\|^2 \leq 2\eta^2\left((1-\beta)^2\|g^r\|^2 + \beta^2\frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i(x^r)\|^2\right).$$

Using Young's inequality, we have for any $q > 0$ that

$$\begin{aligned}\mathbb{E}[\|\nabla f_i(x^r)\|^2] &\leq (1+q)\mathbb{E}[\|\nabla f_i(x^{r-1})\|^2] + (1+q^{-1})L^2\mathbb{E}[\|x^r - x^{r-1}\|^2] \\ &\leq (1+q)\mathbb{E}[\|\nabla f_i(x^{r-1})\|^2] + 2(1+q^{-1})\gamma^2L^2(\mathcal{E}_{r-1} + \mathbb{E}[\|\nabla f(x^{r-1})\|^2]) \\ &\leq (1+q)^r\mathbb{E}[\|\nabla f_i(x^0)\|^2] + \frac{2}{q}\gamma^2L^2\sum_{j=0}^{r-1}(\mathcal{E}_j + \mathbb{E}[\|\nabla f(x^j)\|^2)(1+q)^{r-j}.\end{aligned}$$

By letting $q = \frac{1}{r}$, we have

$$\mathbb{E}[\|\nabla f_i(x^r)\|^2] \leq e\mathbb{E}[\|\nabla f_i(x^0)\|^2] + 2e(r+1)\gamma^2 L^2 \sum_{j=0}^{r-1} (\mathcal{E}_j + \mathbb{E}[\|\nabla f(x^j)\|^2]). \tag{B.1}$$

Note that this inequality is valid for $r = 0$. Therefore, using (B.1), we have

$$\begin{aligned}
\sum_{r=0}^{R-1} \Xi_r &\leq \sum_{r=0}^{R-1} 2\eta^2 \mathbb{E}\left[(1-\beta)^2\|g^r\|^2 + \beta^2 \frac{1}{N}\sum_{i=1}^{N} \|\nabla f_i(x^r)\|^2\right] \\
&\leq \sum_{r=0}^{R-1} 2\eta^2 \left(2(1-\beta)^2(\mathcal{E}_{r-1} + \mathbb{E}[\|\nabla f(x^{r-1})\|^2]) + \beta^2 \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[\|\nabla f_i(x^r)\|^2]\right) \\
&\leq \sum_{r=0}^{R-1} 4\eta^2(1-\beta)^2(\mathcal{E}_{r-1} + \mathbb{E}[\|\nabla f(x^{r-1})\|^2]) \\
&\quad + 2\eta^2\beta^2 \sum_{r=0}^{R-1} \left(\frac{e}{N}\sum_{i=1}^{N} \mathbb{E}[\|\nabla f_i(x^0)\|^2] + 2e(r+1)(\gamma L)^2 \sum_{j=0}^{r-1}(\mathcal{E}_j + \mathbb{E}[\|\nabla f(x^j)\|^2])\right) \\
&\leq 4\eta^2(1-\beta)^2 \sum_{r=0}^{R-1} (\mathcal{E}_{r-1} + \mathbb{E}[\|\nabla f(x^{r-1})\|^2]) \\
&\quad + 2\eta^2\beta^2 \left(eRG_0 + 2e(\gamma LR)^2 \sum_{r=0}^{R-2} (\mathcal{E}_r + \mathbb{E}[\|\nabla f(x^r)\|^2])\right).
\end{aligned}$$

Rearranging the equation and applying upper bound of η completes the proof. □

Theorem B.4. *Under Assumption 1 and 3, if we take* $g^0 = 0$, $\beta = \min\left\{1, \sqrt{\frac{NKL\Delta}{\sigma^2 R}}\right\}$, $\gamma = \min\left\{\frac{1}{24L}, \frac{\beta}{6L}\right\}$, *and* $\eta KL \lesssim \min\left\{1, \frac{1}{\beta\gamma LR}, \left(\frac{L\Delta}{G_0\beta^3 R}\right)^{1/2}, \frac{1}{(\beta N)^{1/2}}, \frac{1}{(\beta^3 NK)^{1/4}}\right\}$, *then* FedAvg-M *converges as*

$$\frac{1}{R}\sum_{r=0}^{R-1} \mathbb{E}[\|\nabla f(x^r)\|^2] \lesssim \sqrt{\frac{L\Delta\sigma^2}{NKR}} + \frac{L\Delta}{R}.$$

Proof. Combining Lemma B.1 and B.2, we have

$$\begin{aligned}
\mathcal{E}_r \leq \left(1 - \frac{8\beta}{9}\right)\mathcal{E}_{r-1} + 4\frac{(\gamma L)^2}{\beta}\mathbb{E}[\|\nabla f(x^{r-1})\|^2] + \frac{2\beta^2\sigma^2}{NK} \\
+ 4\beta L^2\left(2eK^2\Xi_r + K\eta^2\beta^2\sigma^2(1 + 2K^3L^2\eta^2\beta^2\right).
\end{aligned}$$

and

$$\mathcal{E}_0 \leq (1-\beta)\mathcal{E}_{-1} + \frac{2\beta^2\sigma^2}{NK} + 4\beta L^2\left(2eK^2\Xi_0 + K\eta^2\beta^2\sigma^2(1 + 2K^3L^2\eta^2\beta^2)\right).$$

Summing over r from 0 to $R-1$ and applying Lemma B.3, we have

$$\begin{aligned}
\sum_{r=0}^{R-1}\mathcal{E}_r \le{}& \left(1-\frac{8\beta}{9}\right)\sum_{r=-1}^{R-2}\mathcal{E}_r + 4\frac{(\gamma L)^2}{\beta}\sum_{r=0}^{R-2}\mathbb{E}[\|\nabla f(x^r)\|^2] + 2\frac{\beta^2\sigma^2}{NK}R \\
&+ 4\beta L^2\left(2eK^2\sum_{r=0}^{R-1}\Xi_r + RK\eta^2\beta^2\sigma^2(1+2K^3L^2\eta^2\beta^2)\right) \\
\le{}& \left(1-\frac{7\beta}{9}\right)\sum_{r=-1}^{R-2}\mathcal{E}_r + \left(4\frac{(\gamma L)^2}{\beta}+\frac{\beta}{9}\right)\sum_{r=-1}^{R-2}\mathbb{E}[\|\nabla f(x^r)\|^2] + 16\beta^3(e\eta KL)^2RG_0 \\
&+ \frac{2\beta^2\sigma^2}{NK}R + 4\beta^3(\eta KL)^2\left(\frac{1}{K}+2(\eta KL\beta)^2\right)\sigma^2R \\
\le{}& \left(1-\frac{7\beta}{9}\right)\sum_{r=-1}^{R-2}\mathcal{E}_r + \frac{2\beta}{9}\sum_{r=-1}^{R-2}\mathbb{E}[\|\nabla f(x^r)\|^2] + 16\beta^3(e\eta KL)^2RG_0 + \frac{4\beta^2\sigma^2}{NK}R.
\end{aligned}$$

Here in the last inequality we apply

$$4\beta(\eta KL)^2\left(\frac{1}{K}+2(\eta KL\beta)^2\right) \le \frac{2}{NK} \quad \text{and} \quad \gamma L \le \frac{\beta}{6}.$$

Therefore,

$$\sum_{r=0}^{R-1}\mathcal{E}_r \le \frac{9}{7\beta}\mathcal{E}_{-1} + \frac{2}{7}\mathbb{E}[\sum_{r=-1}^{R-2}\|\nabla f(x^r)\|^2] + \frac{144}{7}(e\beta\eta KL)^2G_0R + \frac{36\beta\sigma^2}{7NK}R.$$

Combining this inequality with Lemma A.1, we get

$$\frac{1}{\gamma}\mathbb{E}[f(x^R)-f(x^0)] \le -\frac{1}{7}\sum_{r=0}^{R-1}\mathbb{E}[\|\nabla f(x^r)\|^2] + \frac{39}{56\beta}\mathcal{E}_{-1} + \frac{78}{7}(e\beta\eta KL)^2G_0R + \frac{39\beta\sigma^2}{14NK}R.$$

Finally, noticing that $g^0=0$ implies $\mathcal{E}_{-1}\le 2L(f(x^0)-f^*)=2L\Delta$, we obtain

$$\begin{aligned}
\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}[\|\nabla f(x^r)\|^2] &\lesssim \frac{L\Delta}{\gamma LR} + \frac{\mathcal{E}_{-1}}{\beta R} + (\beta\eta KL)^2G_0 + \frac{\beta\sigma^2}{NK} \\
&\lesssim \frac{L\Delta}{R} + \frac{L\Delta}{\beta R} + \frac{\beta\sigma^2}{NK} + (\beta\eta KL)^2G_0 \\
&\lesssim \frac{L\Delta}{R} + \sqrt{\frac{L\Delta\sigma^2}{NKR}}.
\end{aligned}$$

□

B.2 FedAvg-M-VR

B.2.1 Algorithm

When each local loss function is further assumed to be sample-wise smooth (*i.e.*, Assumption 2), we can replace the local descent direction in Algorithm 1 with a variance-reduced momentum direction

$$g_i^{r,k} = \nabla F(x_i^{r,k};\xi_i^{r,k}) + (1-\beta)(g^r - \nabla F(x^{r-1};\xi_i^{r,k}))$$

to further enhance convergence, leading to FedAvg-M-VR, as presented in Algorithm 3. Here, the variable x^{r-1} is the last-iterate global model maintained in the server. Same as FedAvg-M, turning off the variance-reduced momentum of FedAvg-M-VR, *i.e.*, setting $\beta = 1$, recovers FedAvg. FedAvg-M-VR shares the same algorithmic structure and uplink communication workload as FedAvg.

Algorithm 3 FedAvg-M-VR: FedAvg with variance-reduced momentum

```
Require: initial model $x^{-1} = x^0$ and gradient estimate $g^0$, local learning rate $\eta$, global learning rate $\gamma$,
  momentum $\beta$
  for $r = 0, \cdots, R-1$ do
    for each client $i \in \{1, \dots, N\}$ in parallel do
      Initial local model $x_i^{r,0} = x^r$
      for $k = 0, \cdots, K-1$ do
        Compute direction $g_i^{r,k} = \nabla F(x_i^{r,k}; \xi_i^{r,k}) + (1-\beta)(g^r - \nabla F(x^{r-1}; \xi_i^{r,k}))$
        Update local model $x_i^{r,k+1} = x_i^{r,k} - \eta g_i^{r,k}$
      end for
    end for
    Aggregate local updates $g^{r+1} = \frac{1}{\eta N K} \sum_{i=1}^{N} \left( x^r - x_i^{r,K} \right)$
    Update global model global $x^{r+1} = x^r - \gamma g^{r+1}$
  end for
```

B.2.2 Convergence analysis

Lemma B.5. *If $\gamma L \leq \sqrt{\frac{\beta N K}{54}}$, the following holds for $r \geq 1$:*

$$\mathcal{E}_r \leq (1 - \frac{8\beta}{9})\mathcal{E}_{r-1} + \frac{4}{\beta} L^2 U_r + \frac{3\beta^2\sigma^2}{NK} + \frac{6(\gamma L)^2}{NK} \mathbb{E}[\|\nabla f(x^{r-1})\|^2.$$

Also for $r = 0$, it holds that

$$\mathcal{E}_0 \leq (1 - \beta)\mathcal{E}_{-1} + \frac{4}{\beta} L^2 U_r + \frac{3\beta^2\sigma^2}{NK}.$$

Proof.

$$\begin{aligned}
\mathcal{E}_r = \mathbb{E}&\left[\left\|\frac{1}{NK}\sum_{i,k}\nabla F(x_i^{r,k};\xi_i^{r,k}) + (1-\beta)\left(g^r - \frac{1}{NK}\sum_{i,k}\nabla F(x^{r-1};\xi_i^{r,k})\right) - \nabla f(x^r)\right\|^2\right] \\
=&\mathbb{E}\Bigg[\Bigg\|(1-\beta)(g^r - \nabla f(x^{r-1})) + \frac{1}{NK}\sum_{i,k}\nabla F(x_i^{r,k};\xi_i^{r,k}) - \nabla f(x^r) \\
&+ (1-\beta)\left(\nabla f(x^{r-1}) - \frac{1}{NK}\sum_{i,k}\nabla F(x^{r-1};\xi_i^{r,k})\right)\Bigg\|^2\Bigg] \\
=&(1-\beta)^2\mathcal{E}_{r-1} + \underbrace{2\mathbb{E}\left[\left\langle (1-\beta)(g^r - \nabla f(x^{r-1})), \frac{1}{NK}\sum_{i,k}\nabla f_i(x_i^{r,k}) - \nabla f(x^r)\right\rangle\right]}_{\Lambda_1} \\
&\underbrace{+\,\mathbb{E}\left\|\frac{1}{NK}\sum_{i,k}\nabla F(x_i^{r,k};\xi_i^{r,k}) - \nabla f(x^r) + (1-\beta)\left(\nabla f(x^{r-1}) - \frac{1}{NK}\sum_{i,k}\nabla F(x^{r-1};\xi_i^{r,k})\right)\right\|^2}_{\Lambda_2}.
\end{aligned}$$

By the AM-GM inequality and Assumption 2, we have

$$\Lambda_1 \le \beta(1-\beta)^2\mathcal{E}_{r-1} + \frac{1}{\beta}L^2 U_r.$$

By Assumption 2,

$$\begin{aligned}
\Lambda_2 = \mathbb{E}&\Bigg[\Bigg\|\frac{1}{NK}\sum_{i,k}\left(\nabla F(x_i^{r,k};\xi_i^{r,k}) - \nabla F(x^r;\xi_i^{r,k})\right) + \beta\left(\frac{1}{NK}\sum_{i,k}\nabla F(x^r;\xi_i^{r,k}) - \nabla f(x^r)\right) \\
&+ (1-\beta)\left(\frac{1}{NK}\sum_{i,k}\left(\nabla F(x^r;\xi_i^{r,k}) - \nabla F(x^{r-1};\xi_i^{r,k})\right) - \nabla f(x^r) + \nabla f(x^{r-1})\right)\Bigg\|^2\Bigg] \\
\le&\ 3L^2U_r + 3\frac{\beta^2\sigma^2}{NK} + 3(1-\beta)^2\frac{L^2}{NK}\mathbb{E}[\|x^r - x^{r-1}\|^2.
\end{aligned}$$

Therefore, for $r \ge 1$,

$$\begin{aligned}
\mathcal{E}_r &\le (1-\beta)\mathcal{E}_{r-1} + \frac{4}{\beta}L^2U_r + \frac{3\beta^2\sigma^2}{NK} + 3(1-\beta)^2\frac{L^2}{NK}\mathbb{E}[\|x^r - x^{r-1}\|^2] \\
&\le (1 - \frac{8\beta}{9})\mathcal{E}_{r-1} + \frac{4}{\beta}L^2U_r + \frac{3\beta^2\sigma^2}{NK} + \frac{6(\gamma L)^2}{NK}\mathbb{E}[\|\nabla f(x^{r-1})\|^2].
\end{aligned}$$

The last inequality is derived by $\|x^r - x^{r-1}\|^2 \le 2\gamma^2(\|\nabla f(x^{r-1})\|^2 + \|g^r - \nabla f(x^{r-1})\|^2)$ and $\gamma L \le \sqrt{\frac{\beta NK}{54}}$. Similarly, for $r = 0$, we can obtain

$$\mathcal{E}_0 \le (1-\beta)\mathcal{E}_{-1} + \frac{4}{\beta}L^2U_0 + \frac{3\beta^2\sigma^2}{NK}.$$

□

Lemma B.6. *If $\eta KL \le \frac{1}{4e}$, the following holds:*

$$U_r \le 4eK^2\Xi_r + 8(\eta K)^2(2(\eta KL)^2 + K^{-1})\left(\beta^2\sigma^2 + 2L^2\mathbb{E}[\|x^r - x^{r-1}\|^2]\right).$$

Proof. Note that $\zeta_i^{r,k} = -\eta(\nabla f_i(x_i^{r,k}) + (1-\beta)(g^r - \nabla f_i(x^{r-1}))$. Then we have

$$\begin{aligned}\mathbb{E}[\|\zeta_i^{r,j} - \zeta_i^{r,j-1}\|^2] &\leq \eta^2 L^2 \mathbb{E}[\|x_i^{r,j} - x_i^{r,j-1}\|^2] \\ &= \eta^2 L^2 \left(\mathbb{E}[\|\zeta_i^{r,j-1}\|^2] + \mathbb{E}[\mathrm{Var}[x_i^{r,j} - x_i^{r,j-1} | \mathcal{F}_i^{r,j-1}]]\right).\end{aligned}$$

Here we use bias-variance decomposition and notation $\mathrm{Var}[\cdot|\cdot]$ stands for the conditional variance. Since

$$\begin{aligned}&\mathbb{E}[\mathrm{Var}[x_i^{r,j} - x_i^{r,j-1} | \mathcal{F}_i^{r,j-1}]] \\ =&\eta^2 \mathbb{E}\left[\left\|\nabla F(x_i^{r,j-1}; \xi_i^{r,j-1}) - \nabla f_i(x_i^{r,j-1}) - (1-\beta)\left(\nabla F(x^{r-1}; \xi_i^{r,j-1}) - \nabla f_i(x^{r-1})\right)\right\|^2\right] \\ \leq&\eta^2 \left(2\beta^2\sigma^2 + 2(1-\beta)^2 L^2 \mathbb{E}[\|x^{r-1} - x_i^{r,j-1}\|^2\right),\end{aligned}$$

then

$$\begin{aligned}&\mathbb{E}[\|\zeta_i^{r,j} - \zeta_i^{r,j-1}\|^2] \\ &\quad\leq \eta^2 L^2 \left(\mathbb{E}[\|\zeta_i^{r,j-1}\|^2 + 2\beta^2\eta^2\sigma^2 + 2\eta^2(1-\beta)^2 L^2 \mathbb{E}[\|x^{r-1} - x_i^{r,j-1}\|^2]\right) \\ &\quad\leq \eta^2 L^2 \left(\mathbb{E}[\|\zeta_i^{r,j-1}\|^2] + 2\beta^2\eta^2\sigma^2 + 4\eta^2 L^2 \mathbb{E}[\|x^{r-1} - x^r\|^2 + \|x^r - x_i^{r,j-1}\|^2]\right).\end{aligned}$$

Therefore for any $1 \leq j \leq k-1 \leq K-2$,

$$\begin{aligned}\mathbb{E}\|\zeta_i^{r,j}\|^2 \leq&(1+\frac{1}{k})\mathbb{E}[\|\zeta_i^{r,j-1}\|^2 + (1+k)\mathbb{E}[\|\zeta_i^{r,j} - \zeta_i^{r,j-1}\|^2] \\ \leq&\left(1+\frac{2}{k}\right)\mathbb{E}\|\zeta_i^{r,j-1}\|^2 + (k+1)\eta^2 L^2 \left(2\beta^2\eta^2\sigma^2 + 4\eta^2 L^2 \mathbb{E}[\|x^{r-1} - x^r\|^2 + \|x^r - x_i^{r,j-1}\|^2]\right) \\ \leq&e^2 \mathbb{E}\|\zeta_i^{r,0}\|^2 + 8k^2 L^2 \eta^4 (2\beta^2\sigma^2 + 4L^2\mathbb{E}[\|x^r - x^{r-1}\|^2]) + 4e^2 k(\eta L)^4 \sum_{j'=0}^{j-1} \mathbb{E}[\|x_i^{r,j'} - x^r\|^2].\end{aligned} \tag{B.2}$$

Here the second inequality is by $\eta L \leq \frac{1}{K} \leq \frac{1}{k+1}$. The last inequality is derived by unrolling the recursive bound and using $\left(1+\frac{2}{k}\right)^k \leq e^2$. By Lemma A.3, it holds that

$$\begin{aligned}\mathbb{E}[\|x_i^{r,k} - x^r\|^2] \leq&2\mathbb{E}\left[\left\|\sum_{j=0}^{k-1}\zeta_i^{r,j}\right\|^2\right] + 2\sum_{j=0}^{k-1}\mathbb{E}[\mathrm{Var}[x_i^{r,j+1} - x_i^{r,j} | \mathcal{F}_i^{r,j}]] \\ \leq&2k\sum_{j=0}^{k-1}\mathbb{E}[\|\zeta_i^{r,j}\|^2] + 2\sum_{j=0}^{k-1}\left(2\beta^2\eta^2\sigma^2 + 4\eta^2 L^2 \mathbb{E}[\|x^{r-1} - x^r\|^2 + \|x^r - x_i^{r,j}\|^2]\right).\end{aligned} \tag{B.3}$$

Summing up (B.3) over $k = 0, \ldots, K-1$, using (B.2) and $8(\eta L)^2 K + 8e^2(\eta K L)^4 \leq \frac{1}{2}$ due to the condition on η, we have

$$\frac{1}{2K}\sum_{k=0}^{K-1}\mathbb{E}[\|x_i^{r,k} - x^r\|^2 \leq 2eK^2\mathbb{E}[\|\zeta_i^{r,0}\|^2] + (8(\eta K)^4 L^2 + 4\eta^2 K)\left(\beta^2\sigma^2 + 2L^2\mathbb{E}[\|x^r - x^{r-1}\|^2]\right).$$

This implies

$$U_r \leq 4eK^2\Xi_r + 8(\eta K)^2(2(\eta K L)^2 + K^{-1})\left(\beta^2\sigma^2 + 2L^2\mathbb{E}[\|x^r - x^{r-1}\|^2]\right).$$

□

Lemma B.7. *If* $\gamma L \le \frac{1}{24}$ *and* $288e(\eta K L)^2\left(\frac{289}{72}(1-\beta)^2 + 8e(\gamma\beta L R)^2\right) \le \beta^2$*, then the following holds:*

$$\sum_{r=0}^{R-1} \Xi_r \le \frac{\beta^2}{288eK^2L^2}\sum_{r=-1}^{R-2}(\mathcal{E}_r + \mathbb{E}[\|\nabla f(x^r)\|^2]) + 4\eta^2\beta^2 eRG_0.$$

Proof. Recall that $\zeta_i^{r,0} = -\eta((1-\beta)(g^r - \nabla f_i(x^{r-1})) + \nabla f_i(x^r))$. Consequently, we have

$$\begin{aligned}\|\zeta_i^{r,0}\|^2 &\le 2\eta^2\left((1-\beta)^2\|g^r\|^2 + \|\nabla f_i(x^r) - (1-\beta)\nabla f_i(x^{r-1})\|^2\right)\\ &\le 2\eta^2(1-\beta)^2(1+2(\gamma L)^2)\|g^r\|^2 + 4\eta^2\beta^2\|\nabla f_i(x^r)\|^2\\ &\le \frac{289}{144}\eta^2(1-\beta)^2\|g^r\|^2 + 4\eta^2\beta^2\|\nabla f_i(x^r)\|^2.\end{aligned}$$

Using the AM-GM inequality, we can obtain that for any $q > 0$,

$$\begin{aligned}\mathbb{E}[\|\nabla f_i(x^r)\|^2] &\le (1+q)\mathbb{E}[\|\nabla f_i(x^{r-1})\|^2] + (1+q^{-1})L^2\mathbb{E}\|x^r - x^{r-1}\|^2\\ &\le (1+q)\mathbb{E}[\|\nabla f_i(x^{r-1})\|^2] + 2(1+q^{-1})(\gamma L)^2(\mathcal{E}_{r-1} + \mathbb{E}[\|\nabla f(x^{r-1})\|^2])\\ &\le (1+q)^r\mathbb{E}[\|\nabla f_i(x^0)\|^2] + \frac{2}{q}(\gamma L)^2\sum_{j=0}^{r-1}(\mathcal{E}_j + \mathbb{E}[\|\nabla f(x^j)\|^2])(1+q)^{r-j}.\end{aligned}$$

Taking $q = \frac{1}{r}$ in the inequality above, we have

$$\mathbb{E}[\|\nabla f_i(x^r)\|^2] \le e\mathbb{E}[\|\nabla f_i(x^0)\|^2] + 2e(r+1)(\gamma L)^2\sum_{j=0}^{r-1}(\mathcal{E}_j + \mathbb{E}[\|\nabla f(x^j)\|^2]).$$

This inequality holds as well trivially for $r = 0$. Therefore, we have

$$\begin{aligned}\sum_{r=0}^{R-1}\Xi_r &\le \sum_{r=0}^{R-1}\mathbb{E}\left[\frac{289}{144}\eta^2(1-\beta)^2\|g^r\|^2 + 4\eta^2\beta^2\frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i(x^r)\|^2\right]\\ &\le \sum_{r=0}^{R-1}\frac{289}{72}\eta^2(1-\beta)^2(\mathcal{E}_{r-1} + \mathbb{E}[\|\nabla f(x^{r-1})\|^2])\\ &\quad + 4\eta^2\beta^2\sum_{r=0}^{R-1}\left(\frac{e}{N}\sum_{i=1}^{N}\mathbb{E}[\|\nabla f_i(x^0)\|^2] + 2e(r+1)(\gamma L)^2\sum_{j=0}^{r-1}(\mathcal{E}_j + \mathbb{E}[\|\nabla f(x^j)\|^2])\right)\\ &\le \frac{289}{72}\eta^2(1-\beta)^2\sum_{r=0}^{R-1}(\mathcal{E}_{r-1} + \mathbb{E}[\|\nabla f(x^{r-1})\|^2])\\ &\quad 4\eta^2\beta^2\left(eRG_0 + 2e(\gamma LR)^2\sum_{r=0}^{R-2}(\mathcal{E}_r + \mathbb{E}[\|\nabla f(x^r)\|^2])\right)\\ &\le \frac{\beta^2}{288eK^2L^2}\sum_{r=-1}^{R-2}(\mathcal{E}_r + \mathbb{E}[\|\nabla f(x^r)\|^2]) + 4\eta^2\beta^2 eRG_0.\end{aligned}$$

Here the last inequality is due to the upper bound of η. □

Theorem B.8. *Under Assumption 2 and 3, if we take* $g^0 = \frac{1}{NB}\sum_{i=1}^{N}\sum_{b=1}^{B}\nabla F(x^0;\xi_i^b)$ *with* $\{\xi_i^b\}_{b=1}^B \overset{iid}{\sim} \mathcal{D}_i$ *and* *set* $\beta = \min\left\{1, \left(\frac{NKL^2\Delta^2}{\sigma^4 R^2}\right)^{1/3}\right\}$, $\gamma = \min\left\{\frac{1}{24L}, \sqrt{\frac{\beta NK}{54L^2}}\right\}$, $B = \left\lceil \frac{K}{R\beta^2} \right\rceil$, *and*

$$\eta KL \lesssim \min\left\{\left(\frac{L\Delta}{G_0\gamma LR}\right)^{1/2}, \left(\frac{\beta}{N}\right)^{1/2}, \left(\frac{\beta}{NK}\right)^{1/4}\right\},$$

then FedAvg-M-VR *converges as*

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}[\|\nabla f(x^r)\|^2] \lesssim \left(\frac{L\Delta\sigma}{NKR}\right)^{2/3} + \frac{L\Delta}{R}.$$

Proof. Combining Lemma B.5 and B.6, we have

$$\begin{aligned}
\mathcal{E}_r &\le (1-\frac{8\beta}{9})\mathcal{E}_{r-1} + \frac{(6\gamma L)^2}{NK}\mathbb{E}[\|\nabla f(x^{r-1})\|^2] + \frac{3\beta^2\sigma^2}{NK} \\
&\quad + \frac{4}{\beta}L^2\left(4eK^2\Xi_r + 8(\eta K)^2(2(\eta KL)^2 + K^{-1})(\beta^2\sigma^2 + 2L^2\mathbb{E}[\|x^r - x^{r-1}\|^2])\right) \\
\mathcal{E}_0 &\le (1-\beta)\mathcal{E}_{-1} + \frac{3\beta^2\sigma^2}{NK} + \frac{4}{\beta}L^2\left(4eK^2\Xi_0 + 8(\eta K)^2(2(\eta KL)^2 + K^{-1}))\beta^2\sigma^2\right)
\end{aligned}$$

Summing over r from 0 to $R-1$ and applying Lemma B.7, we get

$$\begin{aligned}
\sum_{r=0}^{R-1}\mathcal{E}_r &\le (1-\frac{8\beta}{9})\sum_{r=-1}^{R-2}\mathcal{E}_r + \frac{6(\gamma L)^2}{NK}\mathbb{E}\left[\sum_{r=0}^{R-2}\|\nabla f(x^r)\|^2\right] + \frac{3\beta^2\sigma^2}{NK}R \\
&\quad + \frac{4}{\beta}L^2\left(4eK^2\sum_{r=0}^{R-1}\Xi_r + 8(\eta K)^2(2(\eta KL)^2 + \frac{1}{K})\left(R\beta^2\sigma^2 + 2L^2\sum_{r=0}^{R-1}\mathbb{E}[\|x^r - x^{r-1}\|^2]\right)\right) \\
&\le (1-\frac{7\beta}{9})\sum_{r=-1}^{R-2}\mathcal{E}_r + \left(\frac{6(\gamma L)^2}{NK} + \frac{\beta}{9}\right)\mathbb{E}[\sum_{r=-1}^{R-2}\|\nabla f(x^r)\|^2] + 64\beta(e\eta KL)^2RG_0 \\
&\quad + \frac{3\beta^2\sigma^2}{NK}R + 32\beta(\eta KL)^2\left(\frac{1}{K} + 2(\eta KL)^2\right)\sigma^2 R \\
&\le (1-\frac{7\beta}{9})\sum_{r=-1}^{R-2}\mathcal{E}_r + \frac{2\beta}{9}\mathbb{E}\left[\sum_{r=-1}^{R-2}\|\nabla f(x^r)\|^2\right] + 64\beta(e\eta KL)^2RG_0 + \frac{4\beta^2\sigma^2}{NK}R.
\end{aligned}$$

Here in the second inequality we apply

$$\begin{cases}
32\beta(\eta KL)^2(\frac{1}{K} + 2(\eta KL)^2) \le \frac{\beta^2}{NK}, \\
\frac{128(\eta KL)^2}{\beta}(\frac{1}{K} + 2(\eta KL)^2)(\gamma L)^2 \le \frac{\beta}{18}, \\
\gamma L \le \sqrt{\frac{\beta NK}{54}}.
\end{cases}$$

Therefore, we obtain

$$\sum_{r=0}^{R-1}\mathcal{E}_r \le \frac{9}{7\beta}\mathcal{E}_{-1} + \frac{2}{7}\mathbb{E}\left[\sum_{r=-1}^{R-2}\|\nabla f(x^r)\|^2\right] + \frac{576}{7}(e\eta KL)^2G_0R + \frac{36\beta\sigma^2}{7NK}R.$$

Combining this inequality with Lemma A.1, we get

$$\frac{1}{\gamma}\mathbb{E}[f(x^R) - f(x^0)] \le -\frac{1}{7}\sum_{r=0}^{R-1}\mathbb{E}[\|\nabla f(x^r)\|^2] + \frac{39}{56\beta}\mathcal{E}_{-1} + \frac{312}{7}(e\eta KL)^2G_0R + \frac{39\beta\sigma^2}{14NK}R.$$

Finally, noticing that $g^0 = \frac{1}{NB_0}\sum_i \sum_{b=1}^{B} \nabla F(x^0; \xi_i^b)$ implies $\mathcal{E}_{-1} \le \frac{\sigma^2}{NB_0} \le \frac{\beta^2\sigma^2 R}{NK}$, we reach

$$\begin{aligned}
\frac{1}{R}\sum_{r=0}^{R-1} \mathbb{E}[\|\nabla f(x^r)\|^2] &\lesssim \frac{L\Delta}{\gamma LR} + \frac{\mathcal{E}_{-1}}{\beta R} + (\eta KL)^2 G_0 + \frac{\beta\sigma^2}{NK} \\
&\lesssim \frac{L\Delta}{\gamma LR} + \frac{\beta\sigma^2}{NK} \\
&\lesssim \frac{L\Delta}{R} + \frac{L\Delta}{\sqrt{\beta NKR}} + \frac{\beta\sigma^2}{NK} \\
&\lesssim \frac{L\Delta}{R} + \left(\frac{L\Delta\sigma}{NKR}\right)^{2/3}
\end{aligned}$$

□

C SCAFFOLD with momentum

C.1 SCAFFOLD-M

Lemma C.1. *If $\gamma L \le \frac{\beta}{12}$, the following holds for $r \ge 1$:*

$$\mathcal{E}_r \le \left(1 - \frac{8\beta}{9}\right)\mathcal{E}_{r-1} + \frac{16}{\beta}(\gamma L)^2 \mathbb{E}[\|\nabla f(x^{r-1})\|^2] + \frac{4\beta^2\sigma^2}{SK} + 10\beta L^2 U_r + 6\beta^2 \frac{N-S}{S(N-1)} V_r.$$

In addition,

$$\mathcal{E}_0 \le (1-\beta)\mathcal{E}_{-1} + \frac{4\beta^2\sigma^2}{SK} + 8\beta L^2 U_0 + 4\beta^2 \frac{N-S}{S(N-1)} V_0.$$

Proof. Note that $\frac{1}{N}\sum_{i=1}^{N} c_i^r = c^r$ holds for any $r \ge 0$. Using Lemma A.4, we have

$$\begin{aligned}
\mathcal{E}_r &= \mathbb{E}\left[\left\|\nabla f(x^r) - \frac{1}{NK}\sum_{i,\,k} g_i^{r,k}\right\|^2\right] + \frac{N-S}{S(N-1)}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\frac{1}{K}\sum_k g_i^{r,k} - \frac{1}{NK}\sum_{j,k} g_j^{r,k}\right\|^2\right] \\
&= \underbrace{\mathbb{E}\left[\left\|(1-\beta)(\nabla f(x^r) - g^r) + \beta\left(\frac{1}{NK}\sum_{i,\,k}\nabla F(x_i^{r,k};\xi_i^{r,k}) - \nabla f(x^r)\right)\right\|^2\right]}_{\Lambda_1} \\
&\quad + \frac{\beta^2(N-S)}{S(N-1)} \underbrace{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\frac{1}{K}\sum_k \nabla F(x_i^{r,k};\xi_i^{r,k}) - \frac{1}{NK}\sum_{j,k}\nabla F(x_j^{r,k};\xi_j^{r,k}) - (c_i^r - c^r)\right\|^2\right]}_{\Lambda_2}.
\end{aligned}$$

For $r \ge 1$, similar to the proof of Lemma B.1, we have

$$\Lambda_1 \le (1-\beta)\mathcal{E}_{r-1} + \frac{2}{\beta}L^2 \mathbb{E}[\|x^r - x^{r-1}\|^2] + \frac{2\beta^2\sigma^2}{NK} + 4\beta L^2 U_r.$$

Besides, by AM-GM inequality and Lemma A.3,

$$
\begin{aligned}
\Lambda_2 &\leq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\frac{1}{K}\sum_{k}\nabla F(x_i^{r,k};\xi_i^{r,k}) - c_i^r\right\|^2\right] \\
&\leq \frac{2\sigma^2}{K} + \frac{2}{N}\sum_{i}\mathbb{E}\left[\left\|\frac{1}{K}\sum_{k}\nabla f_i(x_i^{r,k}) - c_i^r\right\|^2\right] \\
&\leq \frac{2\sigma^2}{K} + 6(L^2U_r + L^2\mathbb{E}[\|x^r - x^{r-1}\|^2] + V_r).
\end{aligned}
$$

Since $\mathbb{E}[\|x^r - x^{r-1}\|^2] \leq 2\gamma^2(\mathcal{E}_{r-1} + \mathbb{E}[\|\nabla f(x^{r-1})\|^2])$ and $\left(\frac{2}{\beta} + 6\beta^2\frac{N-S}{S(N-1)}\right)2(\gamma L)^2 \leq \frac{16}{\beta}(\gamma L)^2 \leq \frac{\beta}{9}$, we have

$$
\mathcal{E}_r \leq \left(1 - \frac{8\beta}{9}\right)\mathcal{E}_{r-1} + \frac{16}{\beta}(\gamma L)^2\mathbb{E}[\|\nabla f(x^{r-1})\|^2] + \frac{4\beta^2\sigma^2}{SK} + 10\beta L^2U_r + 6\beta^2\frac{N-S}{S(N-1)}V_r.
$$

The case for $r = 0$ is similar. □

Lemma C.2. *If $\gamma L \leq \frac{1}{\sqrt{2\beta}}$ and $\eta KL \leq \frac{1}{\beta}$, it holds for all $r \geq 1$ that*

$$
U_r \leq \eta^2K^2\left(8e(\mathcal{E}_{r-1} + 2\mathbb{E}[\|\nabla f(x^{r-1})\|^2] + \beta^2V_r) + \beta^2\sigma^2(K^{-1} + 2(\beta\eta KL)^2)\right).
$$

Proof. Since $\zeta_i^{r,k} = \mathbb{E}[x_i^{r,k+1} - x_i^{r,k}|\mathcal{F}_i^{r,k}] = -\eta(\beta\nabla f_i(x_i^{r,k}) + (1-\beta)g_r - \beta(c_i^r - c^r))$ and $\mathrm{Var}[x_i^{r,k+1} - x_i^{r,k}|\mathcal{F}_i^{r,k}] \leq \beta^2\eta^2\sigma^2$, with exactly the same procedures of Lemma B.2, we have

$$
U_r \leq 2eK^2\Xi_r + K\eta^2\beta^2\sigma^2(1 + 2K^3L^2\eta^2\beta^2).
$$

Additionally, by AM-GM inequality,

$$
\begin{aligned}
\Xi_r &= \frac{\eta^2}{N}\sum_{i}\mathbb{E}[\|\beta\nabla f_i(x^r) + (1-\beta)g^r - \beta(c_i^r - c^r)\|^2] \\
&= \frac{\eta^2}{N}\sum_{i}\mathbb{E}\left[\|\beta(\nabla f_i(x^r) - \nabla f_i(x^{r-1})) + (1-\beta)(g^r - \nabla f(x^{r-1}))\right. \\
&\qquad \left. -\beta\left(c_i^r - c^r - \nabla f_i(x^{r-1}) + \nabla f(x^{r-1})\right) + \nabla f(x^{r-1})\|^2\right] \\
&\leq 4\eta^2\left(\beta^2L^2\mathbb{E}[\|x^r - x^{r-1}\|^2] + (1-\beta)^2\mathcal{E}_{r-1} + \beta^2V_r + \mathbb{E}[\|\nabla f(x^{r-1})\|^2]\right) \\
&\leq 4\eta^2(\mathcal{E}_{r-1} + 2\mathbb{E}[\|\nabla f(x^{r-1})\|^2] + \beta^2V_r).
\end{aligned}
$$

Plug this inequality into the above bound completes the proof. □

Lemma C.3. *Under the same conditions of Lemma C.2, if $\beta\eta KL \leq \frac{1}{24K^{1/4}}$ and $\eta K \leq \frac{N}{5S}\gamma$, then we have*

$$
\sum_{r=0}^{R-1}V_r \leq \frac{3N}{S}\left(V_0 + \frac{4SR}{NK}\sigma^2 + \frac{8N}{S}(\gamma L)^2\sum_{r=-1}^{R-2}(\mathcal{E}_r + \mathbb{E}[\|\nabla f(x^r)\|^2])\right).
$$

Proof. Note that

$$
c_i^{r+1} = \begin{cases} c_i^r & \text{with probability } 1 - \dfrac{S}{N} \\ \dfrac{1}{K}\displaystyle\sum_{k}\nabla F(x_i^{r,k};\xi_i^{r,k}) & \text{with probability } \dfrac{S}{N}. \end{cases}
$$

Using Young's inequality repeatedly, we have

$$
\begin{aligned}
V_{r+1} &= \left(1-\frac{S}{N}\right)\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|c_i^r-\nabla f_i(x^r)\|^2]+\frac{S}{N}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\frac{1}{K}\sum_{k}\nabla F(x_i^{r,k};\xi_i^{r,k})-\nabla f_i(x^r)\right\|^2\right]\\
&\leq \left(1-\frac{S}{N}\right)\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|c_i^r-\nabla f_i(x^r)\|^2]+\frac{S}{N}\left(\frac{2\sigma^2}{K}+2L^2U_r\right)\\
&\leq \left(1-\frac{S}{N}\right)\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left(1+\frac{S}{2N}\right)\|c_i^r-\nabla f_i(x^{r-1})\|^2+\left(1+\frac{2N}{S}\right)L^2\|x^r-x^{r-1}\|^2\right]\\
&\quad+\frac{2S}{N}\left(\frac{\sigma^2}{K}+L^2U_r\right)\\
&\leq \left(1-\frac{S}{2N}\right)V_r+\frac{2N}{S}L^2\mathbb{E}[\|x^r-x^{r-1}\|^2]+\frac{2S\sigma^2}{NK}+\frac{2S}{N}L^2U_r.
\end{aligned}
$$

Here we apply Lemma A.3 to obtain the second inequality. Combining this with Lemma C.2, we get

$$
\begin{aligned}
V_{r+1} &\leq \left(1-\frac{S}{2N}+16e\frac{S}{N}(\beta\eta KL)^2\right)V_r+2\sigma^2\left(\frac{S}{NK}+\frac{2S}{N}(\beta\eta KL)^2(K^{-1}+2(\beta\eta KL)^2)\right)\\
&\quad+\left(\frac{4N}{S}(\gamma L)^2+\frac{32eS}{N}(\eta KL)^2\right)(\mathcal{E}_{r-1}+\mathbb{E}[\|\nabla f(x^{r-1})\|^2])\\
&\leq \left(1-\frac{S}{3N}\right)V_r+\frac{4S}{NK}\sigma^2+\frac{8N}{S}(\gamma L)^2(\mathcal{E}_{r-1}+\mathbb{E}[\|\nabla f(x^{r-1})\|^2]),
\end{aligned}
$$

where we apply the upper bound of η. Therefore, we finish the proof by summing up over r from 0 to $R-1$ and rearranging the inequality. □

Theorem C.4. *Under Assumption 1 and 3, if we take* $g^0=0$, $c_i^0=\frac{1}{B}\sum_{b=1}^{B}\nabla F(x^0;\xi_i^b)$ *with* $\{\xi_i^b\}_{b=1}^{B}\overset{iid}{\sim}\mathcal{D}_i$, $c^0=\frac{1}{N}\sum_{i=1}^{N}c_i^0$ *and set* $B=\left\lceil\frac{NK}{SR}\right\rceil$, $\gamma=\frac{\beta}{L}$, $\beta=\min\left\{1,\frac{S}{N^{2/3}},\sqrt{\frac{L\Delta SK}{\sigma^2R}},\sqrt{\frac{L\Delta S^2}{G_0N}}\right\}$, $\eta KL\lesssim\min\left\{\frac{1}{S^{1/2}},\frac{1}{\beta K^{1/4}},\frac{S^{1/2}}{N}\right\}$, *then* SCAFFOLD-M *converges as*

$$
\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}[\|\nabla f(x^r)\|^2]\lesssim\sqrt{\frac{L\Delta\sigma^2}{SKR}}+\frac{L\Delta}{R}\left(1+\frac{N^{2/3}}{S}\right).
$$

Proof. By Lemma C.1, we can get the following inequality by summing over r from 0 to $R-1$ and plugging in Lemma C.2 and Lemma C.3

$$\begin{aligned}
\sum_{r=0}^{R-1}\mathcal{E}_r &\le \left(1-\frac{8\beta}{9}\right)\sum_{r=-1}^{R-2}\mathcal{E}_r + \frac{16}{\beta}(\gamma L)^2\sum_{r=0}^{R-2}\mathbb{E}[\|\nabla f(x^r)\|^2] \\
&\quad + \frac{4\beta^2\sigma^2}{SK}R + 10\beta L^2\sum_{r=0}^{R-1}U_r + 6\beta^2\frac{N-S}{S(N-1)}\sum_{r=0}^{R-1}V_r \\
&\le \left(1-\frac{8\beta}{9}+80e\beta(\eta KL)^2\right)\sum_{r=-1}^{R-2}\mathcal{E}_r + (\frac{16}{\beta}(\gamma L)^2+160e\beta(\eta KL)^2)\sum_{r=0}^{R-2}\mathbb{E}[\|\nabla f(x^r)\|^2] \\
&\quad + \beta^2\sigma^2 R\left(\frac{4}{SK}+10(\eta KL)^2(K^{-1}+2(\beta\eta KL)^2)\right) + \\
&\quad + \beta^2\left(6\frac{N-S}{S(N-1)}+80e\beta(\eta KL)^2\right)\sum_{r=0}^{R-1}V_r \\
&\le \left(1-\frac{7\beta}{9}\right)\sum_{r=-1}^{R-2}\mathcal{E}_r + \left(\frac{16}{\beta}(\gamma L)^2+\frac{\beta}{9}\right)\sum_{r=0}^{R-2}\mathbb{E}[\|\nabla f(x^r)\|^2] + \frac{80\beta^2\sigma^2}{SK}R + \frac{30\beta^2 N}{S^2}V_0.
\end{aligned}$$

Here the coefficients in the last inequality is derived by the following bounds:

$$\begin{cases}
160e\beta(\eta KL)^2 + 24(\frac{\beta\gamma LN}{S})^2\left(6\frac{N-S}{S(N-1)}+80e\beta(\eta KL)^2\right) \le \frac{\beta}{9}, \\[2ex]
10(\eta KL)^2(K^{-1}+2(\beta\eta KL)^2) + 960e\beta K^{-1}(\eta KL)^2 \le \frac{4}{SK}, \\[2ex]
80e\beta(\eta KL)^2 \le \frac{4}{S},
\end{cases}$$

which can be guaranteed by

$$\begin{cases}
\gamma L \lesssim \frac{S^{3/2}}{\beta^{1/2}N}, \\[2ex]
\eta KL \lesssim \frac{1}{S^{1/2}}.
\end{cases}$$

Therefore,

$$\sum_{r=0}^{R-1}\mathcal{E}_r \le \frac{9}{7\beta}\mathcal{E}_{-1} + \frac{2}{7}\mathbb{E}\left[\sum_{r=-1}^{R-2}\|\nabla f(x^r)\|^2\right] + \frac{270\beta N}{7S^2}V_0 + \frac{720\beta\sigma^2}{7SK}R.$$

Combining this inequality with Lemma A.1, we obtain

$$\frac{1}{\gamma}\mathbb{E}[f(x^R)-f(x^0)] \le -\frac{1}{7}\sum_{r=0}^{R-1}\mathbb{E}[\|\nabla f(x^r)\|^2] + \frac{39}{56\beta}\mathcal{E}_{-1} + \frac{585\beta N}{28S^2}V_0 + \frac{390\beta\sigma^2}{7SK}R.$$

Finally, noticing that $g^0=0$ implies $\mathcal{E}_{-1}\le 2L\Delta$ and $c_i = \frac{1}{B_0}\sum_b \nabla F(x^0;\xi_i^b)$ implies $V_0 \le \frac{\sigma^2}{B_0} \le \frac{SR\sigma^2}{NK}$, we reach

$$\begin{aligned}
\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}[\|\nabla f(x^r)\|^2] &\lesssim \frac{L\Delta}{\gamma LR} + \frac{\mathcal{E}_{-1}}{\beta R} + \frac{\beta N}{S^2R}V_0 + \frac{\beta\sigma^2}{SK} \\
&\lesssim \frac{L\Delta}{\beta R} + \frac{L\Delta}{S^{3/2}R}N\beta^{1/2} + \frac{\beta\sigma^2}{SK} \\
&\lesssim \frac{L\Delta}{R}\left(1+\frac{N^{2/3}}{S}\right) + \sqrt{\frac{L\Delta\sigma^2}{SKR}}.
\end{aligned}$$

□

C.2 SCAFFOLD-M-VR

C.2.1 Algorithm

When each local loss function is further assumed to be sample-wise smooth (*i.e.*, Assumption 2), we can replace the local descent direction in Algorithm 2 with a variance-reduced momentum direction

$$g_i^{r,k} = \nabla F(x_i^{r,k}; \xi_i^{r,k}) - \beta(c_i^r - c^r) + (1-\beta)(g^r - \nabla F(x^{r-1}; \xi_i^{r,k})),$$

resulting in SCAFFOLD-M-VR, as presented in Algorithm 4. Here, the variable x^{r-1} is the last-iterate global model maintained in the server. Same as SCAFFOLD-M, turning off the variance-reduced momentum of SCAFFOLD-M-VR, *i.e.*, setting $\beta = 1$, recovers SCAFFOLD.

Algorithm 4 SCAFFOLD-M-VR: SCAFFOLD with variance-reduced momentum

```
Require: initial model $x^{-1} = x^0$, gradient estimator $g^0$, control variables $\{c_i^0\}_{i=1}^N$ and $c^0$, local learning rate
  $\eta$, global learning rate $\gamma$, momentum $\beta$
  for $r = 0, \cdots, R-1$ do
    Uniformly sample clients $\mathcal{S}_r \subseteq \{1, \cdots, N\}$ with $|\mathcal{S}_r| = S$
    for each client $i \in \mathcal{S}_r$ in parallel do
      Initialize local model $x_i^{r,0} = x^r$
      for $k = 0, \cdots, K-1$ do
        Compute $g_i^{r,k} = \nabla F(x_i^{r,k}; \xi_i^{r,k}) - \beta(c_i^r - c^r) + (1-\beta)(g^r - \nabla F(x^{r-1}; \xi_i^{r,k}))$
        Update local model $x_i^{r,k+1} = x_i^{r,k} - \eta g_i^{r,k}$
      end for
      Update control variable $c_i^{r+1} := \frac{1}{K}\sum_k \nabla F(x_i^{r,k}; \xi_i^{r,k})$ (for $i \notin \mathcal{S}_r$, $c_i^{r+1} = c_i^r$)
    end for
    Aggregate local updates $g^{r+1} = \frac{1}{\eta SK}\sum_{i\in\mathcal{S}_r}\left(x^r - x_i^{r,K}\right)$
    Update global model $x^{r+1} = x^r - \gamma g^{r+1}$
    Update control variable $c^{r+1} = c^r + \frac{1}{N}\sum_{i\in\mathcal{S}_r}(c_i^{r+1} - c_i^r)$
  end for
```

C.2.2 Convergence analysis

Lemma C.5. *If $\gamma L \le \sqrt{\frac{\beta S}{126}}$, then the following holds for $r \ge 1$:*

$$\mathcal{E}_r \le (1 - \frac{8\beta}{9})\mathcal{E}_{r-1} + \frac{14(\gamma L)^2}{S}\mathbb{E}[\|\nabla f(x^{r-1})\|^2] + \frac{8}{\beta}L^2 U_r + \frac{7\beta^2\sigma^2}{SK} + \frac{4(N-S)}{S(N-1)}\beta^2 V_r.$$

In addition,

$$\mathcal{E}_0 \le (1-\beta)\mathcal{E}_{-1} + \frac{8}{\beta}L^2 U_0 + \frac{7\beta^2\sigma^2}{SK} + \frac{4(N-S)}{S(N-1)}\beta^2 V_0.$$

Proof. By Lemma A.3, we have

$$\mathcal{E}_r \le \underbrace{\mathbb{E}\left[\left\|\nabla f(x^r) - \frac{1}{NK}\sum_{i,\,k}\left[\nabla F(x_i^{r,k};\xi_i^{r,k}) + (1-\beta)(g^r - \nabla F(x^{r-1};\xi_i^{r,k}))\right]\right\|^2\right]}_{\Lambda_1}$$

$$+ \frac{N-S}{S(N-1)}\underbrace{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\frac{1}{K}\sum_{k}\left[\nabla F(x_i^{r,k};\xi_i^{r,k}) - (1-\beta)\nabla F(x^{r-1};\xi_i^{r,k})\right] - \beta c_i^r\right\|^2\right]}_{\Lambda_2}.$$

Using the same derivation as Lemma B.5, we can show that

$$\Lambda_1 \le (1-\beta)\mathcal{E}_{r-1} + \frac{4}{\beta}L^2 U_r + 3\frac{\beta^2\sigma^2}{NK} + 3(1-\beta)^2\frac{L^2}{NK}\mathbb{E}[\|x^r - x^{r-1}\|^2].$$

Additionally, by the AM-GM inequality,

$$\begin{aligned}\Lambda_2 \le \frac{1}{N}\sum_{i=1}^{N} 4\mathbb{E}\Bigg[&\left\|\frac{1}{K}\sum_{k}\left[\nabla F(x_i^{r,k};\xi_i^{r,k}) - \nabla F(x^r;\xi_i^{r,k})\right]\right\|^2 \\ &+ \beta^2\left\|\frac{1}{K}\sum_{k}\nabla F(x^r;\xi_i^{r,k}) - \nabla f_i(x^r)\right\|^2 + \beta^2\|\nabla f_i(x^{r-1}) - c_i^r\|^2 \\ &+ \left\|\beta(\nabla f_i(x^r) - \nabla f_i(x^{r-1})) + \frac{1-\beta}{K}\sum_{k}\left[\nabla F(x^r;\xi_i^{r,k}) - \nabla F(x^{r-1};\xi_i^{r,k})\right]\right\|^2\Bigg] \\ \le 4&\left(L^2 U_r + \frac{\beta^2\sigma^2}{K} + \beta^2 V_r + L^2\mathbb{E}[\|x^r - x^{r-1}\|^2]\right).\end{aligned}$$

Further notice that for $r \ge 1$, it holds that $\mathbb{E}[\|x^r - x^{r-1}\|^2 \le 2\gamma^2(\mathcal{E}_{r-1} + \mathbb{E}[\|\nabla f(x^{r-1})\|^2])$ and

$$(\gamma L)^2\left(\frac{8(N-S)}{S(N-1)} + \frac{6(1-\beta)^2}{NK}\right) \le \frac{14(\gamma L)^2}{S} \le \frac{\beta}{9}.$$

Hence we obtain

$$\mathcal{E}_r \le (1 - \frac{8\beta}{9})\mathcal{E}_{r-1} + \frac{14(\gamma L)^2}{S}\mathbb{E}[\|\nabla f(x^{r-1})\|^2 + \frac{8}{\beta}L^2 U_r + \frac{7\beta^2\sigma^2}{SK} + \frac{4(N-S)}{S(N-1)}\beta^2 V_r.$$

The case for $r=0$ can be established similarly. □

Lemma C.6. *If* $\eta K L \le \frac{1}{4e}$, $\eta K \le \frac{\gamma N}{10S}$, *and* $\gamma L \le \frac{1}{24}$, *then it holds that*

$$\sum_{r=0}^{R-1} V_r \le \frac{3N}{S}\left(V_0 + \frac{4SR}{NK}\sigma^2 + \frac{6N}{S}(\gamma L)^2\sum_{r=-1}^{R-2}(\mathcal{E}_r + \mathbb{E}[\|\nabla f(x^r)\|^2])\right).$$

Proof. Note that $\zeta_i^{r,k} = -\eta(\nabla f_i(x_i^{r,k}) + (1-\beta)(g^r - \nabla f_i(x^{r-1})) - \beta(c_i^r - c^r))$, with the same procedures in Lemma B.6, we have

$$U_r \le 4eK^2\Xi_r + 8(\eta K)^2(2(\eta K L)^2 + K^{-1})\left(\beta^2\sigma^2 + 2L^2\mathbb{E}[\|x^r - x^{r-1}\|^2]\right).$$

Additionally, by the AM-GM inequality,

$$\begin{aligned}
\Xi_r &= \frac{\eta^2}{N}\sum_i \mathbb{E}[\|\nabla f_i(x^r) + (1-\beta)(g^r - \nabla f_i(x^{r-1}) - \beta(c_i^r - c^r)\|^2] \\
&= \frac{\eta^2}{N}\sum_i \mathbb{E}\left[\|(\nabla f_i(x^r) - \nabla f_i(x^{r-1})) + (1-\beta)(g^r - \nabla f(x^{r-1}))\right. \\
&\qquad \left. -\beta\left(c_i^r - c^r - \nabla f_i(x^{r-1}) + \nabla f(x^{r-1})\right) + \nabla f(x^{r-1})\|^2\right] \\
&\le 4\eta^2\mathbb{E}\left[L^2\|x^r - x^{r-1}\|^2 + (1-\beta)^2\mathcal{E}_{r-1} + \beta^2 V_r + \|\nabla f(x^{r-1})\|^2\right] \\
&\le 8\eta^2(\mathcal{E}_{r-1} + \mathbb{E}[\|\nabla f(x^{r-1})\|^2] + \beta^2 V_r).
\end{aligned}$$

Hence, by applying $32(2(\eta KL)^2 + K^{-1})(\gamma L)^2 \le 96(\gamma L)^2 \le 2$, we obtain

$$\begin{aligned}
U_r &\le 32e(\eta K)^2(\mathcal{E}_{r-1} + \mathbb{E}[\|\nabla f(x^{r-1})\|^2] + \beta^2 V_r) \\
&\quad + 8(\eta K)^2(2(\eta KL)^2 + K^{-1})\left(\beta^2\sigma^2 + 2L^2\mathbb{E}[\|x^r - x^{r-1}\|^2]\right) \\
&\le 90(\eta K)^2(\mathcal{E}_{r-1} + \mathbb{E}[\|\nabla f(x^{r-1})\|^2] + \beta^2 V_r) + 8(\beta\eta K)^2(2(\eta KL)^2 + K^{-1})\sigma^2.
\end{aligned} \tag{C.1}$$

Also, similar to Lemma C.3, it still holds that

$$V_{r+1} \le \left(1 - \frac{S}{2N}\right)V_r + \frac{2N}{S}L^2\mathbb{E}[\|x^r - x^{r-1}\|^2 + \frac{2S\sigma^2}{NK} + \frac{2S}{N}L^2 U_r.$$

Combine this with the upper bound of U_r,

$$\begin{aligned}
V_{r+1} &\le \left(1 - \frac{S}{2N} + \frac{180(\beta\eta KL)^2 S}{N}\right)V_r + \left(\frac{4N(\gamma L)^2}{S} + \frac{180(\eta KL)^2 S}{N}\right)(\mathcal{E}_{r-1} + \mathbb{E}[\|\nabla f(x^{r-1})\|^2]) \\
&\quad + \sigma^2\left(\frac{2S}{NK} + 8(\beta\eta KL)^2(2(\eta KL)^2 + K^{-1})\right) \\
&\le \left(1 - \frac{S}{3N}\right)V_r + \frac{6N(\gamma L)^2}{S}(\mathcal{E}_{r-1} + \mathbb{E}[\|\nabla f(x^{r-1})\|^2]) + \frac{4S\sigma^2}{NK},
\end{aligned}$$

where we apply the upper bound of η in the last inequality. Iterating the above inequality completes the proof. □

Theorem C.7. *Under Assumption 2 and 3, if we take* $c_i^0 = \frac{1}{B}\sum_{b=1}^{B}\nabla F(x^0;\xi_i^b)$ *with* $\{\xi_i^b\}_{b=1}^B \overset{iid}{\sim} \mathcal{D}_i$, $g^0 = c^0 = \frac{1}{N}\sum_{i=1}^{N} c_i^0$ *and set* $\beta = \min\left\{\frac{S}{N}, \left(\frac{KL\Delta}{\sigma^2 R}\right)^{2/3} S^{1/3}\right\}$, $\gamma = \min\left\{\frac{1}{L}, \frac{\sqrt{\beta S}}{L}\right\}$, $B = \left\lceil \max\left\{\frac{SK}{NR\beta^2}, \frac{NK}{SR}\right\}\right\rceil$, $\eta KL \lesssim \min\left\{\left(\frac{\beta}{S}\right)^{1/2}, \left(\frac{\beta}{SK}\right)^{1/4}\right\}$, *then* SCAFFOLD-M-VR *converges as*

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}[\|\nabla f(x^r)\|^2] \lesssim \left(\frac{L\Delta\sigma}{S\sqrt{K}R}\right)^{2/3} + \frac{L\Delta}{R}\left(1 + \frac{N^{1/2}}{S}\right).$$

Proof. By Lemma C.5, sum over r from 0 to $R-1$ and plug (C.1), Lemma C.6 in,

$$\begin{aligned}
\sum_{r=0}^{R-1}\mathcal{E}_r \le{}& (1-\frac{8\beta}{9})\sum_{r=-1}^{R-2}\mathcal{E}_r + \frac{14(\gamma L)^2}{S}\sum_{r=0}^{R-2}\mathbb{E}[\|\nabla f(x^r)\|^2] + \frac{7\beta^2\sigma^2}{SK}R \\
&+ \frac{8}{\beta}L^2\sum_{r=0}^{R-1}U_r + 4\beta^2\frac{N-S}{S(N-1)}\sum_{r=0}^{R-1}V_r \\
\le{}& (1-\frac{8\beta}{9}+720\frac{(\eta KL)^2}{\beta})\sum_{r=-1}^{R-2}\mathcal{E}_r + (\frac{14(\gamma L)^2}{S}+720\frac{(\eta KL)^2}{\beta})\sum_{r=0}^{R-2}\mathbb{E}[\|\nabla f(x^r)\|^2] \\
&+ \beta^2\sigma^2R\left(\frac{7}{SK}+\frac{64(\eta KL)^2}{\beta}(K^{-1}+2(\eta KL)^2)\right) \\
&+ \beta^2\left(\frac{4(N-S)}{S(N-1)}+720\frac{(\eta KL)^2}{\beta}\right)\sum_{r=0}^{R-1}V_r \\
\le{}& (1-\frac{7\beta}{9})\sum_{r=-1}^{R-2}\mathcal{E}_r + (\frac{14(\gamma L)^2}{S}+\frac{\beta}{9})\sum_{r=0}^{R-2}\mathbb{E}[\|\nabla f(x^r)\|^2] + 60\frac{\beta^2\sigma^2}{SK}R + 15\frac{\beta^2 N}{S^2}V_0.
\end{aligned}$$

Here the coefficients in the last inequality is derived by the following bounds:

$$\begin{cases}
720\dfrac{(\eta KL)^2}{\beta} + 18(\dfrac{\beta\gamma LN}{S})^2\left(4\dfrac{N-S}{S(N-1)}+720\dfrac{(\eta KL)^2}{\beta}\right) \le \dfrac{\beta}{9}, \\[2ex]
64\dfrac{(\eta KL)^2}{\beta}(K^{-1}+2(\eta KL)^2) + 8640\dfrac{(\eta KL)^2}{\beta K} \le \dfrac{5}{SK}, \\[2ex]
720\dfrac{(\eta KL)^2}{\beta} \le \dfrac{1}{S},
\end{cases}$$

which can be guaranteed by

$$\begin{cases}
\gamma L \lesssim \dfrac{S^{3/2}}{\beta^{1/2}N}, \\[2ex]
\eta KL \lesssim \min\{\sqrt{\dfrac{\beta}{S}}, (\dfrac{\beta}{SK})^{1/4}\}.
\end{cases}$$

Therefore, it holds that

$$\sum_{r=0}^{R-1}\mathcal{E}_r \le \frac{9}{7\beta}\mathcal{E}_{-1} + \frac{2}{7}\mathbb{E}\left[\sum_{r=-1}^{R-2}\|\nabla f(x^r)\|^2\right] + \frac{135\beta N}{7S^2}V_0 + \frac{540\beta\sigma^2}{7SK}R.$$

Combining this inequality with Lemma A.1, we get

$$\frac{1}{\gamma}\mathbb{E}[f(x^R)-f(x^0)] \le -\frac{1}{7}\sum_{r=0}^{R-1}\mathbb{E}[\|\nabla f(x^r)\|^2] + \frac{39}{56\beta}\mathcal{E}_{-1} + \frac{585\beta N}{56S^2}V_0 + \frac{585\beta\sigma^2}{14SK}R.$$

Finally, noticing that $g^0 = \dfrac{1}{NB_0}\sum_{i,b}\nabla F(x^0;\xi_i^b)$ implies $\mathcal{E}_{-1} \le \dfrac{\sigma^2}{NB_0} \le \dfrac{\beta^2\sigma^2R}{SK}$ and $c_i = \dfrac{1}{B_0}\sum_b \nabla F(x^0;\xi_i^b)$

implies $V_0 \leq \frac{\sigma^2}{B_0} \leq \frac{SR\sigma^2}{NK}$, we reach

$$\begin{aligned}\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}[\|\nabla f(x^r)\|^2] &\lesssim \frac{L\Delta}{\gamma LR} + \frac{\mathcal{E}_{-1}}{\beta R} + \frac{\beta N}{S^2 R}V_0 + \frac{\beta\sigma^2}{SK} \\ &\lesssim \frac{L\Delta}{R} + \frac{L\Delta}{(\beta S)^{1/2}R} + \frac{L\Delta}{S^{3/2}R}N\beta^{1/2} + \frac{\beta\sigma^2}{SK} \\ &\lesssim \frac{L\Delta}{R}\left(1 + \frac{N^{1/2}}{S}\right) + \left(\frac{L\Delta\sigma}{S\sqrt{K}R}\right)^{2/3}.\end{aligned}$$

□