
A Robust Training Method for Federated Learning with Partial Participation

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Client weighting and partial participation are key techniques in federated learning.
2 They reduce communication costs and maintain a balance in the data used for
3 model training. Numerous strategies are well-established within the research
4 community, leading to growing interest in developing a unified theory. In this
5 paper, we explore this issue in detail. We propose a method that accumulates
6 unused gradients from the current iteration locally and, after full aggregation,
7 leverages them for effective training. Our framework supports a wide class of
8 weighting and sampling heuristics. Furthermore, we show the proposed approach
9 to be robust against clients' periodic disconnection. To validate it, we conduct
10 a series of numerical experiments involving the training of convolutional and
11 transformer-based architectures.

12 1 Introduction

13 Optimization is a cornerstone of training machine learning and neural network models. In a nutshell,
14 almost every AI-based solution aims to minimize an empirical risk [Shalev-Shwartz et al., 2010],
15 which evaluates how well the data is approximated. This process involves adjusting parameters
16 to reduce the discrepancy between predicted outputs and ground truth labels, thereby improving
17 generalization performance. Formally, the problem can be expressed as

$$\min_{x \in \mathbb{R}^d} \left[\frac{1}{n} \sum_{i=1}^n \ell(g(x, a_i), b_i) \right], \quad (1)$$

18 where x denotes the trainable parameters of the model g , (a_i, b_i) is the i -th sample from the dataset
19 with size n , and ℓ is the loss function. Nowadays, there is a variety of methods developed to
20 efficiently solve (1) [Robbins and Monro, 1951, Nesterov, 1983, Kingma and Ba, 2014, Defazio and
21 Mishchenko, 2023]. The current successes of machine/deep learning owe much to the development
22 of powerful numerical techniques that enable training on a huge amount of samples. Large-scale
23 data processing became possible with the advancement of distributed optimization [Verbraeken
24 et al., 2020]. Instead of solving the problem on a single machine, samples are shared among M
25 nodes/devices/clients/machines connected via a server. Hence, the problem (1) transforms into

$$\min_{x \in \mathbb{R}^d} \left[f(x) = \frac{1}{M} \sum_{m=1}^M f_m(x) = \frac{1}{M} \sum_{m=1}^M \frac{1}{n_m} \sum_{i_m=1}^{n_m} \ell(g(x, a_{i_m}), b_{i_m}) \right], \quad (2)$$

26 where n_m is the size of the dataset, stored on m -th device.

27 1.1 Client Weighting

28 Parallel data processing helps to reduce computational time significantly [Zinkevich et al., 2010,
29 Abadi et al., 2016, Joulani et al., 2017]. However, contemporary applications present new challenges.

30 Training samples are often accumulated locally by each specific machine, rather than being collected
 31 and distributed manually. This paradigm with data remaining on edge devices is called federated
 32 learning [Konečný et al., 2016, McMahan et al., 2017, Bonawitz et al., 2019]. In such a setup, local
 33 datasets are typically heterogeneous – they vary in size, distribution, and quality. For instance, one
 34 device may hold unique objects that are poorly represented across the rest of the network, but are
 35 crucial for capturing more dependencies. This leads to the conclusion that some clients may be more
 36 useful than others. Modern approaches usually assign dynamic weights $\{\pi_m\}_{m=1}^M$ and use

$$f(x) = \sum_{m=1}^M \pi_m f_m(x), \text{ s.t. } \pi_m > 0, \sum_{m=1}^M \pi_m = 1 \quad (3)$$

37 to calculate statistics. If the devices are considered to be equivalent, this corresponds to the case
 38 where $\pi_1 = \dots = \pi_M = 1/M$. As a result, more important nodes contribute more significantly to the
 39 global loss. There are many strategies to prioritize the clients known in the literature.

40 **Weighting Based on Data Quality/Quantity.** The most straightforward way to cope with data
 41 imbalance is to consider a number of local samples. McMahan et al. [2017] suggested setting each
 42 coefficient as the constant $\pi_m = n_m/n$. Since then, many modifications of this approach have been
 43 proposed, including federated averaging schemes with momentum [Wang et al., 2019, Reddi et al.,
 44 2020], variance reduction [Liang et al., 2019, Karimireddy et al., 2020] and proximal updates [Li
 45 et al., 2020]. However, this type of weighting ignores heterogeneity in terms of data quality, leading
 46 to bias, e.g. if some client holds an enormous amount of objects with the same labels. To support the
 47 diversity of training samples, Yurochkin et al. [2019] proposed to match the neurons of client neural
 48 networks before averaging. Building on the foundations laid by this work, subsequent works have
 49 explored more efficient approaches extensively [Wang et al., 2020a, Zhang et al., 2022, Yang et al.,
 50 2023, Wu et al., 2023, Kafshgari et al., 2023].

51 **Learned Weighting Strategies.** It is also common to learn weighting strategies instead of using
 52 fixed heuristics. Mohri et al. [2019] were among the first to present results in this direction. They pro-
 53 posed solving the saddle-point problem $\min_{x \in \mathbb{R}^d} \max_{\pi \in \Delta_1^M} \sum_{m=1}^M \pi_m f_m(x)$ to give small weights
 54 to well-trained devices. The idea of optimizing agnostic empirical loss was then generalized by Li
 55 et al. [2019a]. Their q-FedAvg can be reduced to agnostic optimization as one of the special cases.
 56 However, in practice, it is hard to search for appropriate saddle-points [Daskalakis and Panageas,
 57 2018, Jin et al., 2020], especially in federated learning [Sharma et al., 2023]. As a result, the commu-
 58 nity has shifted towards softer adaptive approaches based on local losses [Zhang et al., 2020, Gao
 59 et al., 2022] and gradients [Wang et al., 2020b, Luo et al., 2024].

60 **Robust Weighting.** The idea of assigning weights to the devices found its application in robust
 61 optimization, where malicious clients can disrupt the learning process [Baruch et al., 2019, Xie et al.,
 62 2020, Fang et al., 2020]. To combat such attacks, advanced schemes usually compute $\{\pi_m\}_{m=1}^M$, as
 63 the trust scores of the devices based on their objectives decrease [Xie et al., 2019], local gradients
 64 [Cao et al., 2020, Yan et al., 2023], and the number of local samples [Cao and Lai, 2019]. Recently,
 65 researchers came up with the idea of using a Bayesian approach [Yang et al., 2024].

66 1.2 Client Sampling

67 Another significant issue of federated learning, on par with heterogeneity, is the communication
 68 bottleneck [Tang et al., 2020, Shi et al., 2020]. Sharing information between machines is costly and
 69 can limit the positive effect of parallelism, which is especially tangible when clients send messages
 70 to the server [Kairouz et al., 2021]. This issue is magnified in federated learning, where edge devices
 71 may have unstable network connectivity, and transmitting large updates may be prohibitively slow.
 72 Many techniques exist to reduce communication [Seide et al., 2014, Alistarh et al., 2017, Stich,
 73 2018]. Partial participation is a special one among them [Li et al., 2019b, Yang et al., 2021]. In each
 74 communication round, only a random subset of clients participates in training, while the rest remain
 75 inactive. This approach offloads the server by decreasing the number of updates that need to be
 76 aggregated. Moreover, it provides significant advantages in edge computing, where communication
 77 channels are not equivalent, or some of them may be unavailable. Nowadays, there is a wide range of
 78 heuristics, which allows to choose subset of clients efficiently.

Data-Based Sampling Strategies. Methods from this class rely on zero- and first-order information of local functions. Importance Sampling FedAvg [Rizk et al., 2021] was one of the first such approaches. The authors suggested evaluating the relevance of a device by how large its gradient is relative to the others. Indeed, a small gradient makes a weak contribution to the step. Consequently, communication with this node can be neglected. Nguyen et al. [2020] proposed an orthogonal approach. Their FOLB measures the angle between local and average gradient. If it is negative, then such a device is useless at the current moment. This idea was then developed extensively in [Wu and Wang, 2022, Zhou et al., 2022]. In addition, techniques based on the norms of updates [Chen et al., 2020] and local loss decrease [Cho et al., 2022] were proposed. There are also a number of approaches that dynamically exploit data heterogeneity to maintain balance [Zhang et al., 2023] or support diversity [Chen and Vikalo, 2024].

System-Based Sampling Strategies. Another approach is to use information about the network itself. FedCS [Nishio and Yonetani, 2019] categorizes clients into groups based on their computational power. This strategy saves wall-clock time by avoiding frequent selection of weak devices. Another class of techniques optimizes energy consumption [Xu and Wang, 2020]. Most modern system heterogeneity techniques also incorporate local data considerations [Lai et al., 2021, Li et al., 2022]. F3AST [Ribero et al., 2022] learns an availability-dependent client selection strategy to minimize the impact of variance on the global model’s convergence.

Thus, the community came up with various techniques for weighting and sampling to make partial participation as efficient as possible. The development of each new scheme was challenging in terms of algorithm design and convergence proof. Consequently, a number of papers appeared attempting to propose a theory without utilizing the properties of any particular strategy.

1.3 Unification of Sampling Strategies

Existing papers in this area of research are built around the federated averaging scheme [McMahan et al., 2017]. Li et al. [2019b] proposed an analysis for strongly convex objectives, obtaining a sublinear convergence rate $\mathcal{O}(\kappa^2/K)$, where κ is the condition number. However, they modeled the partial participation environment via unbiased sampling. Cho et al. [2022] were the first to study the unified case with biased devices selection. They derived $\mathcal{O}(\kappa^2/K + \kappa Q)$, where Q is a non-vanishing term that becomes zero solely in the absence of sampling bias. Thus, the authors recovered the results of Li et al. [2019b], but failed to extend the theory to weaker assumptions. The first success in this direction was achieved in [Luo et al., 2022]. This work resolved key questions regarding biased sampling in the strongly convex case. However, the non-convex analysis holds greater significance for applications. For this setting, Wang and Ji [2022] obtained $\mathcal{O}(\sqrt{L}/\sqrt{K} + \delta)$, where L is the smoothness constant and δ is the uniform bound on the difference between local gradients. This result contains the non-vanishing term and does not match the lower bound $\Omega(L/K)$ [Carmon et al., 2020].

Thus, current works in this field rely on FedAvg. As a consequence, their analysis requires boundedness of gradients [Li et al., 2019b, Cho et al., 2022, Luo et al., 2022] or their differences [Wang and Ji, 2022] even in the non-stochastic case. Therefore, there is still no flawless unified theory of partial participation.

1.4 Our Contribution

In contrast to prior works, where partial participation analysis was built upon FedAvg, we introduce our own scheme to leverage client sampling. While existing techniques ignore the information from inactive clients, our approach utilizes it for benefits. Namely, devices accumulate gradient surrogates locally, and the server accounts for them after the full aggregation round. The proposed approach allows weighting and sampling clients according to a variety of strategies, including biased ones. The convergence of our scheme can be proven in both strongly convex and non-convex cases without introducing unnatural assumptions. The obtained rates do not contain non-vanishing terms. To validate the theory, we conduct experiments with RESNET-18 and ViT.

2 Setup

We begin presenting our results with assumptions necessary to prove convergence. First of all, the objective is assumed to be smooth. This requirement is well-established in optimization.

130 **Assumption 1.** The function f is L -smooth, i.e. for all $x, y \in \mathbb{R}^d$ it satisfies

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

131 Neural networks tend to have a complex loss landscape [Cybenko, 1989, Nguyen and Hein, 2018].
 132 Since we are motivated by real-world scenarios, our main goal is to prove convergence in the
 133 non-convex case. For completeness, we also derive results under stronger assumptions.

134 **Assumption 2.** The function f is:

135 (a) **non-convex** with at least one global minimum:

$$\text{there exists } x^* \text{ s.t. } f(x^*) = \inf_{x \in \mathbb{R}^d} f(x) > -\infty.$$

136 (b) μ -strongly convex, i.e. for all $x, y \in \mathbb{R}^d$ it satisfies

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2.$$

137 Federated learning methods usually require a bound on data heterogeneity to provide convergence
 138 guarantees [Khaled et al., 2020, Karimireddy et al., 2020]. In our work, we quantify it via gradients
 139 [Tang et al., 2018, Stich, 2020].

140 **Assumption 3.** Each gradient ∇f_m is similar to the full gradient ∇f , i.e. for all $x \in \mathbb{R}^d$ it satisfies

$$\frac{1}{M} \sum_{m=1}^M \|\nabla f_m(x) - \nabla f(x)\|^2 \leq \delta_1 \|\nabla f(x)\|^2 + \delta_2.$$

141 This assumption is not too strict, since we do not require uniform boundedness ($\delta_1 = 0$). The
 142 following one is imposed to derive convergence of our algorithm with local stochasticity. If one
 143 removes it, our theory still holds.

Assumption 4. Each worker has access to a stochastic gradient $\nabla f_m(x, \xi_m)$. This is an unbiased
 random variable with bounded variance, i.e. for all $x \in \mathbb{R}^d$ it satisfies

$$\begin{aligned} \mathbb{E}_{\xi_m} [\nabla f_m(x, \xi_m)] &= \nabla f_m(x), \\ \mathbb{E}_{\xi_m} [\|\nabla f_m(x, \xi_m) - \nabla f_m(x)\|^2] &\leq \sigma^2. \end{aligned}$$

144 This assumption appears in different forms in a number of classic papers [Stich, 2018, Gower
 145 et al., 2019, Gorbunov et al., 2020]. Next, we consider that weights $\{\pi_m\}_{m=1}^M$ from (3) lie on the
 146 regularized simplex. Namely, $\pi \in \Delta_1^M \cap \left(\bigcap_{m=1}^M \left\{ \pi : e_m^\top \pi + \frac{\alpha}{M} \geq 0 \right\} \right)$, where $1 \leq \alpha \leq M$ is the
 147 regularization parameter and e is the unit basis. This technique is useful for solving a wide range of
 148 tasks [Mehta et al., 2024].

149 3 Algorithms and Analysis

150 3.1 Motivation

151 Existing papers on the unification of client sampling consider FedAvg without any modifications.
 152 Section 1.3 suggests that this approach is not promising due to poor results even under strong
 153 assumptions. A potential direction for future research could be to find a more suitable scheme. Below
 154 we propose an intuition that helps to address this issue.

155 To understand biased sampling, Cho et al. [2022] introduced the definition of selection skew and
 156 utilized it in the analysis. This is exactly the cause of the non-vanishing term in their rate. Indeed,
 157 there is no convergence if, for example, some devices are never selected for communication. However,
 158 we propose that the problem could be solved if we could somehow account for the error accumulated
 159 due to bias. To develop this idea, we formalize the sampling strategy as follows. First, we assign
 160 weights π_m to devices, as described in (3). Next, we define the selection rule of the server as a
 161 stochastic operator $\mathcal{R} : \mathbb{R}^M \rightarrow \mathbb{R}^M$ that zeros some entries of the input vector while retaining the
 162 others. Applying this operator to the introduced vector of weights, it can be seen that the wide variety
 163 of strategies described in Section 1.2 fits this formalism. This applies not only to simple cases of
 164 selecting clients with the highest weights but also to non-trivial ones, such as zeroing the weights of
 165 unavailable nodes.

166 Viewing partial participation as weight vector sparsification reveals connections to well-studied
 167 techniques [Beznosikov et al., 2023]. A state-of-the-art technique to handle it efficiently is error

168 feedback [Stich and Karimireddy, 2020, Richtárik et al., 2021]. Since sampling rules are represented
 169 as compressors, we believe that this idea may be extremely useful in our setting as well. However,
 170 we cannot apply the existing framework directly, as it requires all clients to account for the error at
 171 each algorithm iteration. Error feedback was designed to compress the information, while our goal is
 172 to exclude some clients from an entire epoch.
 173 Thus, we have to address this challenge before proceeding to a unified analysis of partial participation.

174 3.2 Partial Participation without Unavailable Devices

175 To develop the idea proposed in Section 3.1, we present the **Partial Participation with Bias Correction**
 176 framework (PPBC, see Algorithm 1) that supports a wide class of weighting and sampling approaches.
 177 Since computing full-batch gradients is often impractical in modern applications, we also account for
 178 local stochasticity.

Algorithm 1 PPBC

```

1: Input: Start point  $x^{-1,H^{-1}} \in \mathbb{R}^d, g^{-1,H^{-1}} \in \mathbb{R}^d$ , epochs number  $K$ , number of devices  $M$ 
2: Parameters: Stepsize  $\gamma > 0$ , momentum  $0 < \theta < 1$ , regularization  $1 \leq \alpha \leq M$ 
3: for epochs  $k = 0, \dots, K - 1$  do
4:   Initialize  $\pi^k$  // Server weighs clients using any procedure
5:    $\hat{\pi}^k = \hat{\mathcal{R}}^k(\pi^k)$  // Server selects clients to communicate through epoch using any rule  $\hat{\mathcal{R}}$ 
6:    $g_m^{k,0} = 0$  // Each client initializes the gradient surrogate
7:    $x^{k,0} = x^{k-1,H^{k-1}} - \gamma g^{k-1,H^{k-1}}$  // Server initializes the initial point of the epoch
8:   Generate  $H^k \sim \text{Geom}(p)$  // Server generates number of iterations of  $k$ -th epoch
9:   for iterations  $h = 0, \dots, H^k - 1$  do
10:     $\tilde{\pi}^{k,h} = \tilde{\mathcal{R}}^{k,h}(\hat{\pi}^k)$  // Server selects clients to communicate at the current round using rule  $\tilde{\mathcal{R}}$ 
11:    for devices  $m = 1 \dots M$  in parallel do
12:       $g_m^{k,h+1} = g_m^{k,h} + (1 - \theta) \left( \frac{1}{M} - \tilde{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h})$  // Update the gradient surrogate
13:    end for
14:    for each device  $m : \tilde{\pi}_m^{k,h} \neq 0$  do
15:      Send  $\nabla f_m(x^{k,h}, \xi_m^{k,h})$  to the server
16:    end for
17:     $x^{k,h+1} = x^{k,h} - \gamma \left[ (1 - \theta) \sum_{m=1}^M \tilde{\pi}_m^{k,h} \nabla f_m(x^{k,h}, \xi_m^{k,h}) + \theta g^{k-1,H^{k-1}} \right]$  // Server updates
    parameters
18:  end for
19:  for devices  $m = 1 \dots M$  in parallel do
20:    Send  $g_m^{k,H^k}$  to the server
21:  end for
22:   $g^{k,H^k} = \sum_{m=1}^M g_m^{k,H^k}$  // Server aggregates gradient surrogates
23: end for

```

179 **Description of Algorithm 1.** In Algorithm 1, the weights $\pi^k = (\pi_1^k, \dots, \pi_M^k)^\top$ are computed
 180 according to any of the mentioned strategies at the beginning of each epoch (Line 4). Next, the
 181 rule $\hat{\mathcal{R}}$ is applied to determine the participating machines (Line 5). Its output $\hat{\pi}^k$ contains zeros at
 182 positions corresponding to nodes that are not chosen to communicate with the server. Note that $\hat{\mathcal{R}}$ is
 183 not necessarily constant. There are no theoretical restrictions to change it during the execution. For
 184 example, one can vary the number of participating devices. We also allow additional client sampling
 185 at each iteration of the epoch by introducing a rule $\tilde{\mathcal{R}}$ (Line 10). We propose to aggregate local
 186 gradient surrogates during the epoch (Line 12). To provide intuition beyond this update, we give a toy
 187 example where each π_m is equal to $1/M$. In this way, all inactive devices collect their gradients, while
 188 all active ones retain the vector g_m from the previous iteration. In the practical case with various
 189 weights, each device accounts for its deviation from the uniform distribution $\pi_u = \{1/M\}_{m=1}^M$. Next,
 190 we use the accumulated vectors during the following epoch (Line 17). To handle the magnitude
 191 imbalance between the gradient and its surrogate, we employ a smoothing scheme with a small
 192 parameter θ .

193 **Analysis of Algorithm 1.** We utilize virtual sequences to derive convergence rates of PPBC. The
 194 idea is to introduce an additional vector

$$\tilde{x}^{k,h} = x^{k,h} - \gamma \sum_{m=1}^M g_m^{k,h}$$

195 and use it to prove convergence. Substituting Lines 10, 17 in this definition, we obtain

$$\tilde{x}^{k,h+1} = \tilde{x}^{k,h} - \gamma \left[(1-\theta) \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) + \theta g^{k-1, H^{k-1}} \right].$$

196 This is an important technique for our method, since the sequence \tilde{x} is updated with the average
 197 of gradients from all devices, contrary to the original x . However, the virtual update also contains
 198 a combination of accumulated gradients from the previous epoch. We emphasize that handling
 199 $g^{k-1, H^{k-1}}$ is one of the main theoretical challenges we address. We set the epoch size H^k as a
 200 geometrically distributed random variable and provide the following lemma.

201 **Lemma 1.** Suppose Assumptions 3, 4 hold. We consider the epoch size $H^k \sim \text{Geom}(p)$ and
 202 $1 \leq \alpha \leq M$. Then for Algorithm 1 it implies

$$\begin{aligned} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k,0}} \dots \mathbb{E}_{\xi_m^{k, H^k-1}} \|g^{k, H^k}\|^2 &\leq \frac{24(1-\theta)^2 \alpha (\delta_1 + 1)}{p^2} \mathbb{E}_{H^k} \|\nabla f(x^{k, H^k})\|^2 + \frac{48(1-\theta)^2 \alpha \delta_2}{p^2} \\ &\quad + \frac{24(1-\theta)^2 \alpha \sigma^2}{Mp^2}. \end{aligned}$$

203 Assumption 4 is required only to handle local stochasticity. If the devices are able to compute exact
 204 gradients, Lemma 1 holds with $\sigma = 0$. For the details, see Appendix D. As a result, we obtain the
 205 convergence theorem.

206 **Theorem 1.** Suppose Assumptions 1, 2(a), 3, 4 hold. Then for Algorithm 1 with $\theta \leq \frac{\gamma L p^2}{2}$ and
 207 $\gamma \leq \frac{p}{384 L \alpha (\delta_1 + 1)}$ it implies that

$$\begin{aligned} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(x^{k,0})\|^2 &\leq \frac{16(f(x^{0,0}) - f(x^*))}{\gamma K} + \frac{768 \gamma L \alpha \delta_2}{p} + \frac{384 \gamma^2 L^2 \alpha \delta_2}{p^3} \\ &\quad + \frac{400 \gamma L \alpha \sigma^2}{Mp} + \frac{192 \gamma^2 L^2 \alpha \sigma^2}{Mp^3}. \end{aligned}$$

208 The main obstacle in proving Theorem 1 is the terms $\|g^{k, H^k}\|^2$ and $\|g^{k-1, H^{k-1}}\|^2$ that appear in
 209 the analysis. Using Lemma 1, they can be screwed to $\|\nabla f(x^{k, H^k})\|^2$ and $\|\nabla f(x^{k-1, H^{k-1}})\|^2$,
 210 respectively. The first norm is easy to analyze. Classically, it serves as a convergence criterion.
 211 Eliminating the second one turns out to be challenging. To cope with it, we incorporate the surrogate
 212 into the starting point of the epoch (Line 7). For the details, see Appendix D.1. With such an estimate,
 213 there is a technique to choose the stepsize γ appropriately to obtain convergence [Stich, 2019].

214 **Corollary 1.** Under conditions of Theorem 1 Algorithm 1 with fixed rules $\hat{\mathcal{R}}^k \equiv \tilde{\mathcal{R}}^{k,h} \equiv \mathcal{R}$ needs

$$\mathcal{O} \left(M \frac{M}{C} \left(\frac{\Delta L \alpha \delta_1}{\varepsilon^2} + \frac{\Delta L \alpha \delta_2}{\varepsilon^4} + \frac{\Delta L \alpha \sigma^2}{M \varepsilon^4} \right) \right)$$

215 number of devices communications to reach ε -accuracy, where $\varepsilon^2 = \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(x^{k,0})\|^2$, $\Delta =$
 216 $f(x^{0,0}) - f(x^*)$ and C is the number of devices participating in each epoch.

217 We also consider varying sampling rules $\hat{\mathcal{R}}^k$ and $\tilde{\mathcal{R}}^{k,h}$ to study corollaries of Theorem 1, see
 218 Appendix D.1 for the details. In our work, the analysis is extended to the strongly convex case.

219 **Theorem 2.** Suppose Assumptions 1, 2(b), 3, 4 hold. Then for Algorithm 1 with $\theta \leq \frac{p \gamma \mu}{4}$ and
 220 $\gamma \leq \frac{p^2}{96 L \alpha (\delta_1 + 1)}$ it implies that

$$\mathbb{E} \|x^{K,0} - x^*\|^2 \leq \left(1 - \frac{\gamma \mu}{8}\right)^K \|x^{0,0} - x^*\|^2 + \frac{8 \gamma \alpha}{\mu p^3} \left(144 \delta_2 + \frac{74 \sigma^2}{M}\right).$$

221 As well as for the non-convex objective, suitable γ can be chosen in Theorem 2.

222 **Corollary 2.** Under conditions of Theorem 2 Algorithm 1 with fixed rules $\widehat{\mathcal{R}}^{k,h} \equiv \widetilde{\mathcal{R}}^{k,h} \equiv \mathcal{R}$ needs

$$\widetilde{\mathcal{O}} \left(M \left(\frac{M}{C} \right)^2 \left(\frac{L}{\mu} \alpha \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{C} \frac{\alpha \delta_2}{\mu^2 \varepsilon} + \frac{\alpha \sigma^2}{\mu^2 C \varepsilon} \right) \right)$$

223 number of devices communications to reach ε -accuracy, where $\varepsilon^2 = \mathbb{E} \|x^{K,0} - x^*\|^2$ and C is the
224 number of devices participating in each epoch.

225 3.3 Partial Participation with Unavailable Devices

226 The previous section addresses partial participation when all devices are available to communicate
227 with the server. Indeed, in Algorithm 1 each node receives the current parameters at the end of the
228 iteration, but does not send its gradient. This is motivated by the fact that forwarding a message
229 from the client to the server is much more expensive than the other way around [Kairouz et al.,
230 2021]. However, in practice, some devices can become inactive periodically [Li et al., 2019b, Yang
231 et al., 2021]. Namely, these machines not only refrain from transmitting information but also do not
232 perform local computations. In this section, we extend our theory to cover the case where the actual
233 parameters are sent to only a fraction of the clients.

234 **Description of Algorithm 2.** In this section we present the part of Algorithm 2 (see Appendix A)
235 that reflects key differences from Algorithm 1. To design it, we refuse using the biased sampling
236 rule $\widetilde{\mathcal{R}}$ during the epoch. Instead, we simulate outage probability of the m -th device as a Bernoulli
237 random variable $\eta_m^{k,h} \sim \text{Be}(q_m)$ [Chung, 2000] (Line 9). To describe client disconnection formally,
238 $\eta_m^{k,h}$ is used to update the gradient surrogates (Line 11) and to perform the step (Line 16). Thus, in
239 practice, it is not necessary for an inactive device to know the actual parameters. We also normalize
240 the computed gradients by factors $\{q_m\}_{m=1}^M$ to balance their magnitudes.

9: Generate $\eta^{k,h}$

11: $g_m^{k,h+1} = g_m^{k,h} + (1 - \theta) \frac{\eta_m^{k,h}}{q_m} \left(\frac{1}{M} - \hat{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h})$

16: $x^{k,h+1} = x^{k,h} - \gamma \left[(1 - \theta) \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \hat{\pi}_m^{k,h} \nabla f_m(x^{k,h}, \xi_m^{k,h}) + \theta g^{k-1, H^{k-1}} \right]$

241 **Analysis of Algorithm 2.** We formulate the results for both non-convex and strongly-convex cases.

242 **Corollary 3.** Suppose Assumptions 1, 2(a), 3, 4 hold. Algorithm 2 with fixed rules $\widehat{\mathcal{R}}^k \equiv \widetilde{\mathcal{R}}^{k,h} \equiv \mathcal{R}$
243 needs

$$\mathcal{O} \left(M \frac{1}{C \min_{1 \leq m \leq M} q_m} \left(\frac{\Delta L \alpha \delta_1}{\varepsilon^2} + \frac{\Delta L \alpha \delta_2}{\varepsilon^4} + \frac{\Delta L \alpha \sigma^2}{\varepsilon^4} \right) \right)$$

244 number of devices communications to reach ε -accuracy, where $\varepsilon^2 = \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(x^{k,0})\|^2$, $\Delta =$
245 $f(x^{0,0}) - f(x^*)$ and C is the number of devices participating in each epoch.

246 **Corollary 4.** Suppose Assumptions 1, 2(b), 3, 4 hold. Algorithm 2 with fixed rules $\widehat{\mathcal{R}}^k \equiv \widetilde{\mathcal{R}}^{k,h} \equiv \mathcal{R}$
247 needs

$$\widetilde{\mathcal{O}} \left(M \left(\frac{M}{C} \right)^2 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{L}{\mu} \alpha \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{C} \frac{\alpha \delta_2}{\mu^2 \varepsilon} + \frac{M}{C} \frac{\alpha \sigma^2}{\mu^2 \varepsilon} \right) \right)$$

248 number of devices communications to reach ε -accuracy, where $\varepsilon^2 = \mathbb{E} \|x^{K,0} - x^*\|^2$ and C is the
249 number of devices participating in each epoch.

250 For more details, see Appendix E. Note that $\min_{1 \leq m \leq M} q_m$ is a constant lying in the interval $(0, 1]$.
251 Thus, the rates of Algorithm 2 do not differ significantly from those for Algorithm 1. The only
252 deterioration occurs in the variance term associated with local stochasticity. Thus, if each device has
253 an access to its exact gradient, there is no asymptotical difference compared to Corollaries 1 and 2.

254 3.4 Discussion

255 We analyzed a wide class of sampling and weighting techniques and proposed algorithms for different
256 network scenarios. Their rates asymptotically coincide with the optimal ones for SGD-like approaches

[Stich, 2019]. Due to considering biased strategies, we obtained an additional factor M/C . Again analogizing to compression, this multiplier signifies compression power. It is a well-known fact that there is no theoretical improvement for methods built upon error-feedback [Richtárik et al., 2021, Beznosikov et al., 2023]. However, we recover the convergence of SGD in the case of full participation. Comparing our non-convex rate regarding the main term $\mathcal{O}(1/\varepsilon^2)$ with prior works, we note that it surpasses that in [Wang and Ji, 2022] ($\mathcal{O}(1/\varepsilon^4 + \delta_2)$) both asymptotically and by the absence of the non-vanishing term. Next, comparing strongly-convex rates ($\mathcal{O}(\kappa \log 1/\varepsilon)$), we are superior to [Cho et al., 2022] ($\mathcal{O}(\kappa^2/\varepsilon + \kappa\delta_2)$) and [Luo et al., 2022] ($\mathcal{O}(\kappa/\varepsilon)$). Moreover, both of these works lack non-convex analysis. We highlight that we soften assumptions from all aforementioned works.

4 Experiments

To validate our theoretical findings, we conduct a systematic empirical study comparing three optimization approaches — FedAvg [Reddi et al., 2020], SCAFFOLD [Karimireddy et al., 2020], and PPBC (Algorithm 1) — each integrated with the same client sampling technique. Crucially, we maintain a fixed strategy across all methods, deliberately decoupling the sampling mechanism from algorithmic innovations to focus specifically on its interaction with different optimization approaches. Our experiments assess their relative performance under identical conditions, including model architectures, benchmark datasets, and hardware configurations. Firstly, we detail the experimental setup, including the neural network architectures, benchmark datasets, and computational hardware configurations employed in our analysis.

Experimental Setup. We evaluate each sampling strategy under three distinct data distribution scenarios: **(distr-1)** homogeneous (*i.i.d.*), **(distr-2)** heterogeneous with different classes on each client, and **(distr-3)** strongly heterogeneous configurations with varying client-specific data quantities and class distributions. Our benchmark experiments employ image classification on CIFAR-10 [Krizhevsky et al., 2009] using a RESNET-18 architecture [Meng et al., 2019], establishing a controlled testbed for comparative analysis of Algorithm 1 across the sampling strategies. Comprehensive details regarding data partitioning, model architecture, and dataset specifications are provided in Appendix B.

Client Selection Rule. Notably, not all strategies included in our comparative analysis inherently incorporate a client selection mechanism. To ensure a fair and consistent evaluation, we uniformly applied the following selection rule across all methods:

$$\widehat{\mathcal{R}}^k = \text{Top}_C(\pi^k),$$

where Top_C denotes taking $C > 0$ clients with the highest weights π^k . Consequently, the remainder of our experiments will focus exclusively on the formulation and analysis of weight update rules, while treating the client selection process itself as a fixed component of the experimental framework.

Loss-aware client sampling. Building upon previous work, Cho et al. [2022] introduced the POWER-OF-CHOICE (PoC) strategy, which employs a weighted client sampling mechanism based on local loss values. Formally, the weight update rule can be expressed as:

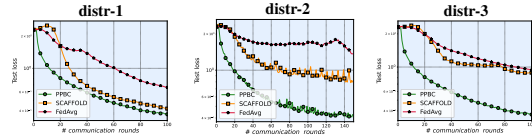
1. The server assigns to all clients the probabilities proportional to the data size fractions

$$p_m = \frac{n_m}{\left(\sum_{m'=1}^M n_{m'}\right)}.$$

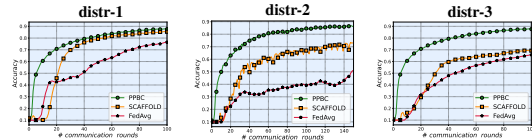
2. The global model is sent by the server to the selected C clients, which compute and return their local loss values based on their datasets. Subsequently, the weights are updated:

$$\pi^k = \left(\left[\frac{1}{n_m} \sum_{i_m=1}^{n_m} \ell(g(x, a_{i_m}), b_{i_m}) \right] \right)_{m=1}^M.$$

Trust-Score Sampling. The study by Xie et al. [2019] introduces the BANT, which implements



(a) Convergence comparison.



(b) Metrics comparison.

Figure 1: Performance comparison for PoC strategy with different data distributions.

a trust-based sampling mechanism. This approach assigns dynamic trust scores to clients based on historical performance metrics. Thus, weight update rule can be described as:

1. The server assigns trust scores TS_m^k to each client m based on the alignment of their model updates with the performance on server-held ground truth data \mathcal{V} :

$$\text{TS}_m^k = \exp \left[-\frac{1}{|\mathcal{V}|} \sum_{\xi \in \mathcal{V}} f_m(x^k, \xi) \right].$$

2. The weights are updated with a probability proportional to the trust scores assigned to each client:

$$\pi^k = \left(\frac{\text{TS}_m^k}{\sum_{m'=1}^M \text{TS}_{m'}^k} \right)_{m=1}^M.$$

Importance Sampling. Nguyen et al. [2020] introduced FOLB, a theoretically grounded client selection framework for federated learning that optimizes convergence by sampling clients proportionally to the expected utility of their local updates. The core selection mechanism operates as follows:

1. Each client is assigned an importance score IS_m^k proportional to the inner product between its gradient $\nabla f_m(x^k, \xi_m^k)$ and the direction of the server model improvement (previous gradient d^k):

$$\text{IS}_m^k = |\langle \nabla f_m(x^k, \xi_m^k), d^k \rangle|.$$

2. The weights are updated with a probability proportional to the trust scores for each client:

$$\pi^k = \left(\frac{\text{IS}_m^k}{\sum_{m'=1}^M \text{IS}_{m'}^k} \right)_{m=1}^M.$$

Discussion. Our experimental evaluation on the CIFAR-10 dataset using the RESNET18 architecture demonstrates a substantial performance gap between conventional approaches (FedAvg, SCAFFOLD) and Algorithm 1 (see Figures 1, 2, and 3), providing strong empirical validation of our theoretical analysis. Notably, PPBC maintains consistent convergence rates and accuracy across all experimental configurations, with the observed performance variance remaining within 2% of theoretical predictions (see Figure 4). This robust empirical behavior confirms our key theoretical insight: PPBC’s performance is strategy-agnostic, achieving stable convergence regardless of the underlying client selection mechanism.

We present additional experiments in Appendix

B. We consider advanced client selection techniques that utilize the rule $\tilde{\mathcal{R}}^{k,h}$, provide the results for PPBC+ (Algorithm 2), and demonstrate the outcomes for ViT training.

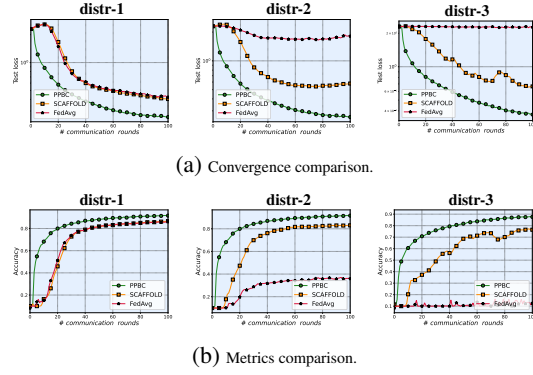


Figure 2: Performance comparison for BANT strategy with different data distributions.

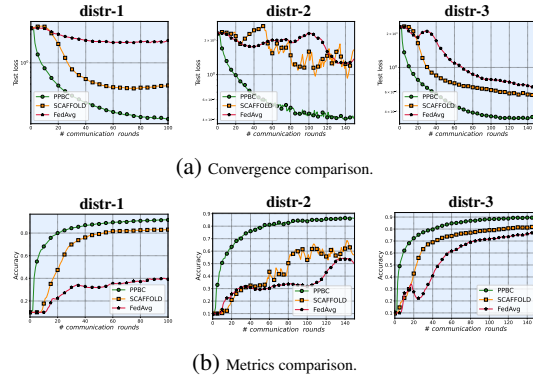


Figure 3: Performance comparison for FOLB strategy with different data distributions.

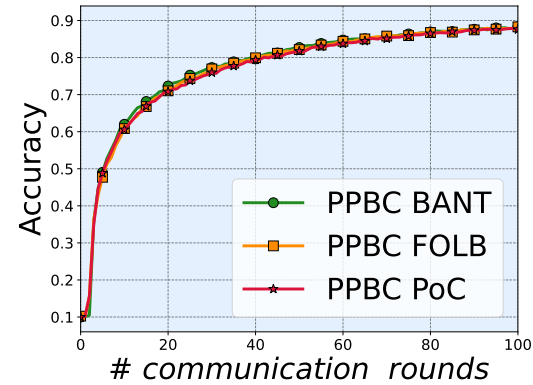


Figure 4: PPBC performance across all client-sampling strategies on the **distr-3**.

References

- Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. {TensorFlow}: a system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283, 2016.
- Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in neural information processing systems*, 30, 2017.
- Zeyuan Allen-Zhu. Katyusha x: Practical momentum method for stochastic sum-of-nonconvex optimization. *arXiv preprint arXiv:1802.03866*, 2018.
- Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems*, 32, 2019.
- Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *Journal of Machine Learning Research*, 24(276):1–50, 2023.
- Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečný, Stefano Mazzocchi, Brendan McMahan, et al. Towards federated learning at scale: System design. *Proceedings of machine learning and systems*, 1: 374–388, 2019.
- Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.
- Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. Fltrust: Byzantine-robust federated learning via trust bootstrapping. *arXiv preprint arXiv:2012.13995*, 2020.
- Xinyang Cao and Lifeng Lai. Distributed gradient descent algorithm robust to an arbitrary number of byzantine attackers. *IEEE Transactions on Signal Processing*, 67(22):5850–5864, 2019.
- Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1):71–120, 2020. doi:[10.1007/s10107-019-01406-y](https://doi.org/10.1007/s10107-019-01406-y). URL <https://doi.org/10.1007/s10107-019-01406-y>.
- Huancheng Chen and Haris Vikalo. Heterogeneity-guided client sampling: Towards fast and efficient non-iid federated learning. *Advances in Neural Information Processing Systems*, 37:65525–65561, 2024.
- Wenlin Chen, Samuel Horvath, and Peter Richtarik. Optimal client sampling for federated learning. *arXiv preprint arXiv:2010.13723*, 2020.
- Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Towards understanding biased client selection in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 10351–10375. PMLR, 2022.
- Kai Lai Chung. *A course in probability theory*. Elsevier, 2000.
- George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.
- Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. *Advances in neural information processing systems*, 31, 2018.
- Aaron Defazio and Konstantin Mishchenko. Learning-rate-free learning by d-adaptation. In *International Conference on Machine Learning*, pages 7449–7479. PMLR, 2023.
- Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. Local model poisoning attacks to {Byzantine-Robust} federated learning. In *29th USENIX security symposium (USENIX Security 20)*, pages 1605–1622, 2020.

384 Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated
385 learning with non-iid data via local drift decoupling and correction. In *Proceedings of the*
386 *IEEE/CVF conference on computer vision and pattern recognition*, pages 10112–10121, 2022.

387 Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of sgd: Variance reduc-
388 tion, sampling, quantization and coordinate descent. In *International Conference on Artificial*
389 *Intelligence and Statistics*, pages 680–690. PMLR, 2020.

390 Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter
391 Richtárik. Sgd: General analysis and improved rates. In *International conference on machine*
392 *learning*, pages 5200–5209. PMLR, 2019.

393 Ali Hatamizadeh, Greg Heinrich, Hongxu Yin, Andrew Tao, Jose M Alvarez, Jan Kautz, and
394 Pavlo Molchanov. Fastervit: Fast vision transformers with hierarchical attention. *arXiv preprint*
395 *arXiv:2306.06189*, 2023.

396 Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave
397 minimax optimization? In *International conference on machine learning*, pages 4880–4889.
398 PMLR, 2020.

399 Norman P Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa,
400 Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, et al. In-datacenter performance analysis of
401 a tensor processing unit. In *Proceedings of the 44th annual international symposium on computer*
402 *architecture*, pages 1–12, 2017.

403 Zahra Hafezi Kafshgari, Chamani Shiranthika, Parvaneh Saeedi, and Ivan V Bajić. Quality-adaptive
404 split-federated learning for segmenting medical images with inaccurate annotations. In *2023 IEEE*
405 *20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.

406 Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin
407 Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Ad-
408 vances and open problems in federated learning. *Foundations and trends® in machine learning*,
409 14(1–2):1–210, 2021.

410 Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and
411 Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In
412 *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

413 Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical
414 and heterogeneous data. In *International conference on artificial intelligence and statistics*, pages
415 4519–4529. PMLR, 2020.

416 Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint*
417 *arXiv:1412.6980*, 2014.

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

421 Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

422 Fan Lai, Xiangfeng Zhu, Harsha V Madhyastha, and Mosharaf Chowdhury. Oort: Efficient federated
423 learning via guided participant selection. In *15th {USENIX} Symposium on Operating Systems*
424 *Design and Implementation ({OSDI} 21)*, pages 19–35, 2021.

425 Chenning Li, Xiao Zeng, Mi Zhang, and Zhichao Cao. Pyramidfl: A fine-grained client selection
426 framework for efficient federated learning. In *Proceedings of the 28th annual international*
427 *conference on mobile computing and networking*, pages 158–171, 2022.

428 Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated
429 learning. *arXiv preprint arXiv:1905.10497*, 2019a.

430 Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith.
431 Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*,
432 2:429–450, 2020.

433 Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of
434 fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019b.

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

437 Bing Luo, Wenli Xiao, Shiqiang Wang, Jianwei Huang, and Leandros Tassiulas. Tackling system and
438 statistical heterogeneity for federated learning with adaptive client sampling. In *IEEE INFOCOM*
439 *2022-IEEE conference on computer communications*, pages 1739–1748. IEEE, 2022.

440 Ping Luo, Xiaoge Deng, Ziqing Wen, Tao Sun, and Dongsheng Li. Accelerating federated learning
441 by selecting beneficial herd of local gradients. *arXiv preprint arXiv:2403.16557*, 2024.

442 Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas.
443 Communication-efficient learning of deep networks from decentralized data. In *Artificial intelli-*
444 *gence and statistics*, pages 1273–1282. PMLR, 2017.

445 Ronak Mehta, Jelena Diakonikolas, and Zaid Harchaoui. Drago: Primal-dual coupled variance
446 reduction for faster distributionally robust optimization. In *The Thirty-eighth Annual Conference*
447 *on Neural Information Processing Systems*, 2024.

448 Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. Frame attention networks for facial expression
449 recognition in videos. In *2019 IEEE international conference on image processing (ICIP)*, pages
450 3866–3870. IEEE, 2019.

451 Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *Interna-*
452 *tional conference on machine learning*, pages 4615–4625. PMLR, 2019.

453 Yurii Nesterov. A method for solving the convex programming problem with convergence rate o
454 $(1/k^2)$. In *Dokl akad nauk Sssr*, volume 269, page 543, 1983.

455 Hung T Nguyen, Vikash Sehwal, Seyyedali Hosseinalipour, Christopher G Brinton, Mung Chiang,
456 and H Vincent Poor. Fast-convergent federated learning. *IEEE Journal on Selected Areas in*
457 *Communications*, 39(1):201–218, 2020.

458 Quynh Nguyen and Matthias Hein. Optimization landscape and expressivity of deep cnns. In
459 *International conference on machine learning*, pages 3730–3739. PMLR, 2018.

460 Takayuki Nishio and Ryo Yonetani. Client selection for federated learning with heterogeneous
461 resources in mobile edge. In *ICC 2019-2019 IEEE international conference on communications*
462 *(ICC)*, pages 1–7. IEEE, 2019.

463 Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito,
464 Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in
465 pytorch. 2017.

466 Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný,
467 Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint*
468 *arXiv:2003.00295*, 2020.

469 Mónica Ribero, Haris Vikalo, and Gustavo De Veciana. Federated learning under intermittent client
470 availability and time-varying communication constraints. *IEEE Journal of Selected Topics in*
471 *Signal Processing*, 17(1):98–111, 2022.

472 Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. Ef21: A new, simpler, theoretically better,
473 and practically faster error feedback. *Advances in Neural Information Processing Systems*, 34:
474 4384–4396, 2021.

475 Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for
476 the masses, 2021.

477 Elsa Ritzk, Stefan Vlaski, and Ali H Sayed. Optimal importance sampling for federated learning. In
478 *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing*
479 *(ICASSP)*, pages 3095–3099. IEEE, 2021.

480 Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical*
481 *statistics*, pages 400–407, 1951.

482 Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and
483 its application to data-parallel distributed training of speech dnns. In *Interspeech*, volume 2014,
484 pages 1058–1062. Singapore, 2014.

485 Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability
486 and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

487 Pranay Sharma, Rohan Panda, and Gauri Joshi. Federated minimax optimization with client hetero-
488 geneity. *arXiv preprint arXiv:2302.04249*, 2023.

489 Shaohuai Shi, Zhenheng Tang, Xiaowen Chu, Chengjian Liu, Wei Wang, and Bo Li. A quantitative
490 survey of communication optimizations in distributed deep learning. *IEEE Network*, 35(3):230–237,
491 2020.

492 Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*,
493 2018.

494 Sebastian U Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint*
495 *arXiv:1907.04232*, 2019.

496 Sebastian U Stich. On communication compression for distributed optimization on heterogeneous
497 data. *arXiv preprint arXiv:2009.02388*, 2020.

498 Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Sgd with delayed
499 gradients. *Journal of Machine Learning Research*, 21(237):1–36, 2020.

500 Hanlin Tang, Shaoduo Gan, Ce Zhang, Tong Zhang, and Ji Liu. Communication compression for
501 decentralized training. *Advances in Neural Information Processing Systems*, 31, 2018.

502 Zhenheng Tang, Shaohuai Shi, Wei Wang, Bo Li, and Xiaowen Chu. Communication-efficient
503 distributed deep learning: A comprehensive survey. *arXiv preprint arXiv:2003.06307*, 2020.

504 Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S
505 Rellermeyer. A survey on distributed machine learning. *Acm computing surveys (csur)*, 53(2):
506 1–33, 2020.

507 Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni.
508 Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*, 2020a.

509 Jianyu Wang, Vinayak Tania, Nicolas Ballas, and Michael Rabbat. Slowmo: Improving
510 communication-efficient distributed sgd with slow momentum. *arXiv preprint arXiv:1910.00643*,
511 2019.

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

515 Shiqiang Wang and Mingyue Ji. A unified analysis of federated learning with arbitrary client
516 participation. *Advances in Neural Information Processing Systems*, 35:19124–19137, 2022.

517 Chenrui Wu, Zexi Li, Fangxin Wang, and Chao Wu. Learning cautiously in federated learning with
518 noisy and heterogeneous clients. In *2023 IEEE International Conference on Multimedia and Expo*
519 *(ICME)*, pages 660–665. IEEE, 2023.

520 Hongda Wu and Ping Wang. Node selection toward faster convergence for federated learning on
521 non-iid data. *IEEE Transactions on Network Science and Engineering*, 9(5):3099–3111, 2022.

522 Cong Xie, Sanmi Koyejo, and Indranil Gupta. Zeno: Distributed stochastic gradient descent with
523 suspicion-based fault-tolerance. In *International Conference on Machine Learning*, pages 6893–
524 6901. PMLR, 2019.

525 Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Fall of empires: Breaking byzantine-tolerant sgd
526 by inner product manipulation. In *Uncertainty in Artificial Intelligence*, pages 261–270. PMLR,
527 2020.

528 Jie Xu and Heqiang Wang. Client selection and bandwidth allocation in wireless federated learning
529 networks: A long-term perspective. *IEEE Transactions on Wireless Communications*, 20(2):
530 1188–1200, 2020.

531 Haonan Yan, Wenjing Zhang, Qian Chen, Xiaoguang Li, Wenhai Sun, Hui Li, and Xiaodong Lin.
532 Recess vaccine for federated learning: Proactive defense against model poisoning attacks. *Advances*
533 *in Neural Information Processing Systems*, 36:8702–8713, 2023.

534 Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation
535 in non-iid federated learning. *Advances in Neural Information Processing Systems (NeurIPS)*,
536 34:5974–5986, 2021. URL <https://proceedings.neurips.cc/paper/2021/hash/...>
537 NeurIPS 2021.

538 Mingkun Yang, Ran Zhu, Qing Wang, and Jie Yang. Fedtrans: Client-transparent utility estimation for
539 robust federated learning. In *The Twelfth International Conference on Learning Representations*,
540 2024.

541 Zhiqin Yang, Yonggang Zhang, Yu Zheng, Xinmei Tian, Hao Peng, Tongliang Liu, and Bo Han.
542 Fedfed: Feature distillation against data heterogeneity in federated learning. *Advances in Neural*
543 *Information Processing Systems*, 36:60397–60428, 2023.

544 Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and
545 Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International*
546 *conference on machine learning*, pages 7252–7261. PMLR, 2019.

547 Jianyi Zhang, Ang Li, Minxue Tang, Jingwei Sun, Xiang Chen, Fan Zhang, Changyou Chen, Yiran
548 Chen, and Hai Li. Fed-cbs: A heterogeneity-aware client sampling mechanism for federated
549 learning via class-imbalance reduction. In *International Conference on Machine Learning*, pages
550 41354–41381. PMLR, 2023.

551 Lin Zhang, Li Shen, Liang Ding, Dacheng Tao, and Ling-Yu Duan. Fine-tuning global model via
552 data-free knowledge distillation for non-iid federated learning. In *Proceedings of the IEEE/CVF*
553 *conference on computer vision and pattern recognition*, pages 10174–10183, 2022.

554 Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M Alvarez. Personalized federated
555 learning with first order model optimization. *arXiv preprint arXiv:2012.08565*, 2020.

556 Pengyuan Zhou, Hengwei Xu, Lik Hang Lee, Pei Fang, and Pan Hui. Are you left out? an efficient
557 and fair federated learning for personalized profiles on wearable devices of inferior networking
558 conditions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*,
559 6(2):1–25, 2022.

560 Martin Zinkevich, Markus Weimer, Lihong Li, and Alex Smola. Parallelized stochastic gradient
561 descent. *Advances in neural information processing systems*, 23, 2010.

NeurIPS Paper Checklist

1. Claims

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

Answer: [\[Yes\]](#)

Justification: see Sections [1.4](#), [3](#).

Guidelines:

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

2. Limitations

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

Answer: [\[Yes\]](#)

Justification: all the assumptions are introduced in Section [2](#), further limitations are discussed in Section [3.4](#).

Guidelines:

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

3. Theory assumptions and proofs

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

Answer: [\[Yes\]](#)

Justification: see Sections 2, 3, D, E.

Guidelines:

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

4. Experimental result reproducibility

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

Answer: [Yes]

Justification: all necessary descriptions to understand the results of the experiments are provided. See Sections 4, B.

Guidelines:

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

5. Open access to data and code

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

Answer: [Yes]

Justification: see Section B.

Guidelines:

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

6. Experimental setting/details

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

Answer: [Yes]

Justification: see Sections 4, B.

Guidelines:

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

7. Experiment statistical significance

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

Answer: [No]

Justification: since the experiments are more of a theoretical verification, we have no statistical effects associated with running the experiments.

Guidelines:

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

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

8. Experiments compute resources

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

Answer: [Yes]

Justification: see Section B.

Guidelines:

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

9. Code of ethics

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

Answer: [Yes]

Justification: the paper follows the NeurIPS Code of Ethics.

Guidelines:

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

10. Broader impacts

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

Answer: [NA]

Justification: there is no societal impact of the work performed.

Guidelines:

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

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

11. Safeguards

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

Answer: [NA]

Justification: the paper poses no such risks.

Guidelines:

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

12. Licenses for existing assets

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

Answer: [Yes]

Justification: see Sections 4, B.

Guidelines:

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

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

13. New assets

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

Answer: [NA]

Justification: Justification: the paper does not release new assets.

Guidelines:

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

14. Crowdsourcing and research with human subjects

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

Answer: [NA]

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

Guidelines:

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

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

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

Answer: [NA]

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

Guidelines:

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

16. Declaration of LLM usage

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

876 Answer: [NA]

877 Justification: the core method development in this paper does not involve LLMs.

878 Guidelines:

- 879 • The answer NA means that the core method development in this research does not
880 involve LLMs as any important, original, or non-standard components.
- 881 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
882 for what should or should not be described.

883 **Appendix**

884 **Contents**

885	1 Introduction	1
886	1.1 Client Weighting	1
887	1.2 Client Sampling	2
888	1.3 Unification of Sampling Strategies	3
889	1.4 Our Contribution	3
890	2 Setup	3
891	3 Algorithms and Analysis	4
892	3.1 Motivation	4
893	3.2 Partial Participation without Unavailable Devices	5
894	3.3 Partial Participation with Unavailable Devices	7
895	3.4 Discussion	7
896	4 Experiments	8
897	A Partial Participation with Unavailable Devices	23
898	B Additional experiments and details	23
899	C General statements	25
900	D Proofs for Algorithm 1	26
901	D.1 Proof for non-convex case	29
902	D.2 Proof for strongly-convex case	36
903	E Proofs for Algorithm 2	44
904	E.1 Proof for non-convex setting	46
905	E.2 Proof for strongly-convex setting	53

A Partial Participation with Unavailable Devices

In this section, we present Algorithm 2, which is the complete version of algorithm from Section 3.3. This method can be applied to environments where devices do not perform local computations periodically.

Algorithm 2 PPBC+

```

1: Input: Start point  $x^{-1,H^{-1}} \in \mathbb{R}^d, g^{-1,H^{-1}} \in \mathbb{R}^d$ , epochs number  $K$ , number of devices  $M$ 
2: Parameters: Stepsize  $\gamma > 0$ , momentum  $0 < \theta < 1$ , regularization  $1 \leq \alpha \leq M$ 
3: for epochs  $k = 0, \dots, K - 1$  do
4:   Initialize  $\pi^k$  // Server weighs clients using any procedure
5:    $\hat{\pi}^k = \hat{\mathcal{R}}^k(\pi^k)$  // Server selects clients to communicate through epoch using any rule  $\hat{\mathcal{R}}$ 
6:    $g_m^{k,0} = 0$  // Each client initializes the gradient surrogate
7:    $x^{k,0} = x^{k-1,H^{k-1}} - \gamma g^{k-1,H^{k-1}}$  // Server initializes the initial point of the epoch
8:   Generate  $H^k \sim \text{Geom}(p)$  // Server generates number of iterations of  $k$ -th epoch
9:   for iterations  $h = 0, \dots, H^k - 1$  do
10:    for devices  $m = 1 \dots M$  in parallel do
11:      Generate  $\eta_m^{k,h} \sim \mathcal{B}(q_m)$  // Device generates its state: available / unavailable
12:       $g_m^{k,h+1} = g_m^{k,h} + (1 - \theta) \frac{\eta_m^{k,h}}{q_m} \left( \frac{1}{M} - \hat{\pi}_m^k \right) \nabla f_m(x^{k,h}, \xi_m^{k,h})$  // Update the gradient
        surrogate
13:    end for
14:    for each device  $m : \eta_m^{k,h} \neq 0$  and  $\hat{\pi}_m^k \neq 0$  do
15:      Send  $\frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h})$  to the server
16:    end for
17:     $x^{k,h+1} = x^{k,h} - \gamma \left[ (1 - \theta) \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \hat{\pi}_m^k \nabla f_m(x^{k,h}, \xi_m^{k,h}) + \theta g^{k-1,H^{k-1}} \right]$  // Server up-
        dates parameters
18:  end for
19:  for devices  $m = 1 \dots M$  in parallel do
20:    Send  $g_m^{k,H^k}$  to the server
21:  end for
22:   $g^{k,H^k} = \sum_{m=1}^M g_m^{k,H^k}$  // Server aggregates gradient surrogates
23: end for

```

B Additional experiments and details

Our code is available at https://anonymous.4open.science/r/EF25_NIPS-AD4E/.

Hardware Details. The experiments were conducted using Python with the PyTorch deep learning framework [Paszke et al., 2017]. The computational hardware consisted of a server equipped with an Intel Xeon Gold 6342 CPU and two NVIDIA A100 40GB GPUs. The total runtime for all experimental evaluations amounted to approximately 80 hours. To simulate a federated learning environment, data was distributed across clients based on a heterogeneity parameter.

Gradient-Norm-Based Sampling. For the image classification problem on CIFAR-10 dataset, we introduce an alternative client sampling strategy based on gradient norm sampling GNS Wang et al. [2020b], which prioritizes clients whose local updates exhibit larger magnitudes. In particular:

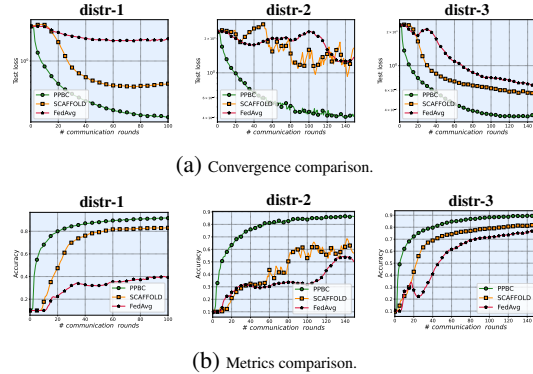


Figure 5: Performance comparison for GNS strategy with different data distributions.

Table 1: Summary of training strategies used in additional experiments. Top and Rand denote the client selection rules, where the number indicates how many clients were selected for training.

Epoch Strategy	Round Strategy
GNS (Top 3)	PoC (Top 1)
FOLB (Top 3)	PoC (Top 1)
PoC (Top 3)	Rand 1
FOLB (Top 3)	Rand 1

1. At each communication round k , the server estimates the relative importance of each client m using the norm of its reported gradient $\nabla f_m(w^k, \xi_m^k)$:

$$p_m^k = \frac{\|\nabla f_m(w^k, \xi_m^k)\|_2}{\sum_{m'=1}^M \|\nabla f_{m'}(w^k, \xi_{m'}^k)\|_2}.$$

2. Clients are then sampled with probabilities proportional to $\{p_m^k\}_{m=1}^M$, ensuring that those with larger gradient norms are selected more frequently:

$$\pi^k = (p_m^k)_{m=1}^M.$$

The obtained comparison results are presented in Figure 5.

ViT Fine-tuning. To further assess the generalization and adaptability of our method, we conduct additional experiments involving the fine-tuning of a state-of-the-art Vision Transformer architecture FASTERViT [Hatamizadeh et al., 2023]. The model, pre-trained on the large-scale IMAGENET21K dataset [Ridnik et al., 2021], comprises approximately 270M parameters and integrates hybrid hierarchical-attention mechanisms for efficient multi-scale feature learning. We fine-tune this model on the FOOD101 dataset [Bossard et al., 2014], a challenging benchmark consisting of 101,000 images across 101 fine-grained food categories. This dataset presents significant visual complexity due to high class variation and subtle inter-class distinctions, making it particularly suitable for evaluating the scalability of our method.

Strategy Mixture. In the preceding experimental setups, we restricted our evaluation to a fixed, server-based client sampling strategy. However, as demonstrated in our theoretical analysis, Algorithm 1 is flexible enough to accommodate a broader class of sampling mechanisms, potentially varying across communication rounds. To validate this flexibility empirically, we conduct additional experiments for FASTERViT fine-tuning on **distr-3** data distribution. We consider this setup to be the most challenging one, because strong heterogeneity with different amount of samples and classes per client and various strategies makes the FedAvg and SCAFFOLD algorithms behave similarly. Therefore, our further experimental comparisons will only include FedAvg. We allow the sampling rule $\widehat{\mathcal{R}}^{k,h}$ to change dynamically at each communication round k . The combinations of strategies are presented in Table 1. The performance validation results for each strategy mixture can be observed in Figure 6.

Unavailable Devices. We now analyze the performance of Algorithm 2, with a focus on scenarios involving *aperiodic client participation*, where certain devices may become temporarily unavailable during training and do not compute error. In this setting, we employ the GNS strategy at the server side, motivated by its strong empirical performance demonstrated by the baseline method FedAvg under the **distr-3** data partitioning scheme. Furthermore, we vary the client sampling parameter q_m , simulating different levels of device availability across communication rounds. The results, summarized in Figure 7, demonstrate that the proposed method, PPBC+,

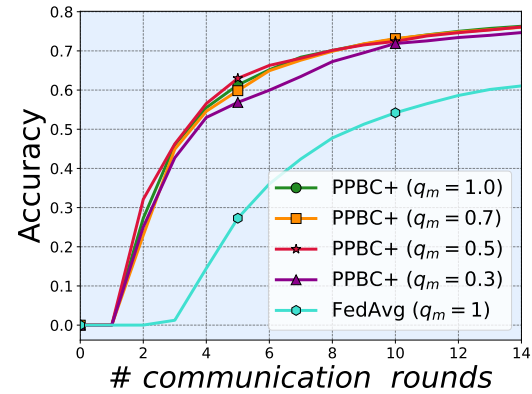


Figure 7: PPBC+ performance for GNS strategy across different q_m values on the **distr-3**.

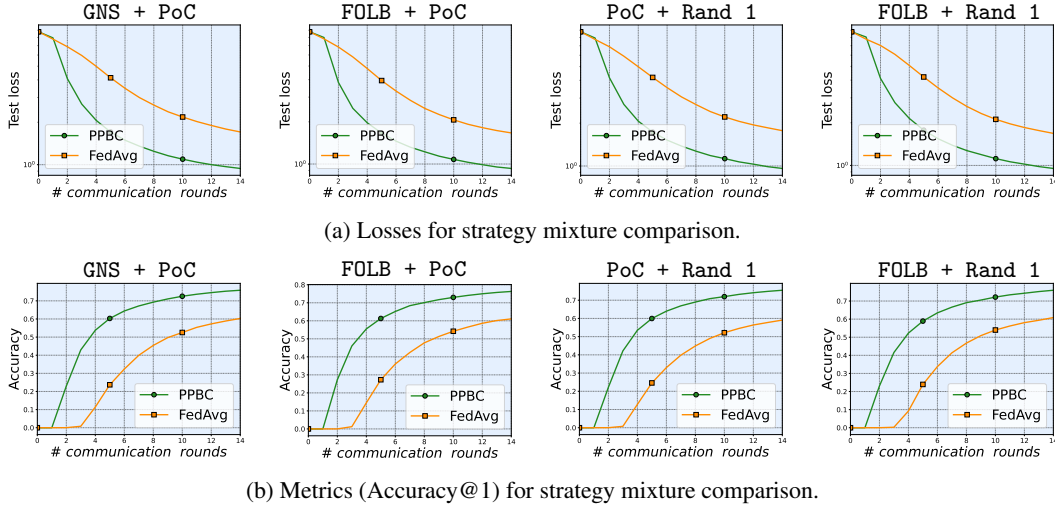


Figure 6: Performance comparison for combination of strategies on FASTERVIT fine-tuning.

maintains superior convergence behavior even under heterogeneous and dynamic participation patterns.

Discussion. We provided experimental validation of the theoretical convergence estimates for the proposed algorithms across a range of practical federated learning tasks. Our evaluation included large-scale models, such as the FASTERVIT architecture with 270M parameters, demonstrating the scalability and effectiveness of our approach in realistic vision-based learning scenarios. Additionally, we analyzed the behavior of the PPBC+ algorithm under varying client sampling conditions, confirming the robustness and consistency of its performance across different parameter q_m values.

To further support our theoretical findings, we present Figure 8, which illustrates that the algorithms introduced in this work maintain comparable convergence rates across all considered configurations. These results affirm that our methods preserve efficiency and stability even when applied to heterogeneous data distributions and complex model architectures.

C General statements

Notation. In the work we use the following notation. $x^{k,h} \in \mathbb{R}^d$ is the vector of model's parameters in h -th iteration in k -th epoch, $\nabla f_m(x) \in \mathbb{R}^d$ represents the gradient of function f_m at the point $x \in \mathbb{R}^d$, $\nabla f_m(x, \xi) \in \mathbb{R}^d$ denotes the stochastic gradient at the point $x \in \mathbb{R}^d$ with respect to stochastic realization ξ .

For a random vector $x \in \mathbb{R}^d$ and stochasticity ξ we denote $\mathbb{E}[x]$ is the expected value of x and $\mathbb{E}_\xi[x]$ as the conditioned expected value with the respect to ξ .

We use $\|x\| = \sqrt{\sum_{i=1}^d x_i^2}$ as l_2 -norm of the vector $x \in \mathbb{R}^d$ and $\langle x, y \rangle = \sum_{i=1}^d x_i y_i$ represents the scalar product of vectors $x, y \in \mathbb{R}^d$.

We use *number of devices communications* (device to server communications) as the metric. This choice arises from the recognition that the number of rounds of communication is insufficient to adequately compare distributed methods. For example, this limitation becomes evident when the nodes operate asynchronously. In this case, the more appropriate metric is the total number of communications rather than the number of rounds.

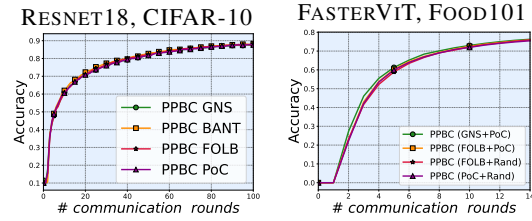


Figure 8: Test accuracy of PPBC for image classification with RESNET18 on CIFAR-10 and FASTERVIT fine-tuning on FOOD101.

1000 **General inequalities.** Suppose $x, y, \{a_i\}_{i=1}^n \in \mathbb{R}^d, \{\omega_i\}_{i=1}^n \in \mathbb{R}, f(\cdot)$ inherent to Assumptions 1,
 1001 2(b), $\varphi(\cdot)$ is under Assumption 2(b). Then,

$$\|\nabla f(x) - \nabla f(y)\|^2 \leq 2L(f(x) - f(y) - \langle \nabla f(y), x - y \rangle), \quad (\text{Lip})$$

$$\langle x, y \rangle \leq \frac{\beta}{2} \|x\|^2 + \frac{1}{2\beta} \|y\|^2, \quad (\text{Fen})$$

$$\left\| \sum_{i=1}^n a_i \right\|^2 \leq n \sum_{i=1}^n \|a_i\|^2, \quad (\text{CS})$$

$$\varphi\left(\frac{\sum_{i=1}^n w_i a_i}{\sum_{i=1}^n a_i}\right) \leq \frac{\sum_{i=1}^n a_i \varphi(x_i)}{\sum_{i=1}^n a_i}. \quad (\text{Jen})$$

1002 **Lemma 2** ([Allen-Zhu, 2018]). Given sequence $D_0, D_1, \dots, D_N \in \mathbb{R}$, where $N \in \text{Geom}(p)$. Then,
 $\mathbb{E}_N[D_{N-1}] = pD_0 + (1-p)\mathbb{E}_N[D_N]$.

1003 D Proofs for Algorithm 1

1004 **Lemma 3 (Lemma 1).** Suppose Assumptions 3, 4 hold. Then for Algorithm 1 it implies that

$$\begin{aligned} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|g^{k,H^k}\|^2 &\leq \frac{24(1-\theta)^2 \alpha(\delta_1 + 1)}{p^2} \mathbb{E}_{H^k} \|\nabla f(x^{k,H^k})\|^2 + \frac{48(1-\theta)^2 \alpha \delta_2}{p^2} \\ &\quad + \frac{24(1-\theta)^2 \alpha \sigma^2}{Mp^2}. \end{aligned}$$

1005 *Proof.* Let us start with the following estimate:

$$\begin{aligned} \|g^{k,h+1}\|^2 &= \left\| g^{k,h} + (1-\theta) \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\ &\stackrel{(\text{Fen})}{\leq} (1+c) \|g^{k,h}\|^2 \\ &\quad + \left(1 + \frac{1}{c}\right) (1-\theta)^2 \left\| \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2, \quad (4) \end{aligned}$$

1006 where c is defined below. Let us estimate the last term and obtain

$$\begin{aligned} &\left\| \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\ &\stackrel{(\text{CS})}{\leq} 2 \left\| \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}) \right\|^2 \\ &\quad + 2 \left\| \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) [\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})] \right\|^2 \\ &\stackrel{(i)}{=} 2 \left\| \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}) - \sum_{m=1}^M \left(\frac{1}{M} - \pi_m^k \right) \nabla f(x^{k,h}) \right\|^2 \\ &\quad + 2 \left\| \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) [\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})] \right\|^2. \end{aligned}$$

1007 Adding and subtracting $\sum_{m=1}^M \tilde{\pi}_m^{k,h} \nabla f(x^{k,h})$ in the first term yields

$$\left\| \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2$$

$$\begin{aligned}
&\leq 2 \left\| \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) [\nabla f_m(x^{k,h}) - \nabla f(x^{k,h})] - \sum_{m=1}^M (\tilde{\pi}_m^{k,h} - \pi_m^k) \nabla f(x^{k,h}) \right\|^2 \\
&\quad + 2 \left\| \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) [\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})] \right\|^2 \\
&\stackrel{\text{(CS)}}{\leq} 4 \left\| \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) [\nabla f_m(x^{k,h}) - \nabla f(x^{k,h})] \right\|^2 \\
&\quad + 4 \left\| \sum_{m=1}^M (\tilde{\pi}_m^{k,h} - \pi_m^k) \nabla f(x^{k,h}) \right\|^2 \\
&\quad + 2 \left\| \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) [\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})] \right\|^2.
\end{aligned}$$

1008 We apply (CS) to the first term and identically transform the second and third terms:

$$\begin{aligned}
&\left\| \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\
&\leq 4 \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right)^2 \sum_{m=1}^M \|\nabla f_m(x^{k,h}) - \nabla f(x^{k,h})\|^2 \\
&\quad + 4 \left(\sum_{m=1}^M (\tilde{\pi}_m^{k,h} - \pi_m^k) \right)^2 \|\nabla f(x^{k,h})\|^2 \\
&\quad + 2 \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right)^2 \|\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})\|^2 \\
&\quad + 4 \sum_{i \neq j} \left(\frac{1}{M} - \tilde{\pi}_i^{k,h} \right) \left(\frac{1}{M} - \tilde{\pi}_j^{k,h} \right) \\
&\quad \cdot \left\langle \nabla f_i(x^{k,h}, \xi_i^{k,h}) - \nabla f_i(x^{k,h}), \nabla f_j(x^{k,h}, \xi_j^{k,h}) - \nabla f_j(x^{k,h}) \right\rangle \\
&\stackrel{\text{As. 3 (ii)}}{\leq} 4M \left(\delta_1 \|\nabla f(x^{k,h})\|^2 + \delta_2 \right) \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right)^2 + 4 \|\nabla f(x^{k,h})\|^2 \\
&\quad + 2 \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right)^2 \|\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})\|^2 \\
&\quad + 4 \sum_{i \neq j} \left(\frac{1}{M} - \tilde{\pi}_i^{k,h} \right) \left(\frac{1}{M} - \tilde{\pi}_j^{k,h} \right) \\
&\quad \cdot \left\langle \nabla f_i(x^{k,h}, \xi_i^{k,h}) - \nabla f_i(x^{k,h}), \nabla f_j(x^{k,h}, \xi_j^{k,h}) - \nabla f_j(x^{k,h}) \right\rangle,
\end{aligned}$$

1009 where (i) was made due to $\sum_{m=1}^M \left(\frac{1}{M} - \pi_m^k \right) = 1 - 1 = 0$, (ii) with respect to $\sum_{m=1}^M (\tilde{\pi}_m^{k,h} - \pi_m^k) \leq 1$.

1010 Taking expectation on $\xi_m^{k,h}$ and using Assumption 4, we have

$$\begin{aligned}
\mathbb{E}_{\xi_m^{k,h}} \left\| \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 &\leq 4M\delta_1 \|\nabla f(x^{k,h})\|^2 \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right)^2 \\
&\quad + 4 \|\nabla f(x^{k,h})\|^2
\end{aligned}$$

$$\begin{aligned}
& +4M\delta_2 \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right)^2 \\
& +2\sigma^2 \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right)^2, \tag{5}
\end{aligned}$$

1011 since $\xi_i^{k,h}$ and $\xi_j^{k,h}$ are independent random variables and, consequently, the scalar product equals to
1012 zero.

1013 We use $\pi^k \in \Delta_1^M \cap \left(\bigcap_{m=1}^M \{ \pi : e_m^\top \pi + \frac{\alpha}{M} \geq 0 \} \right)$, where $1 \leq \alpha \leq M$ and $\{e_m\}_{m=1}^M$ is the unit
1014 basis. In this way, worst case in terms of average distance from $\frac{1}{M}$ is realization, where $\lfloor \frac{M}{\alpha} \rfloor$ weights
1015 are $\frac{\alpha}{M}$ and the rest are zero. In such a case, we can estimate

$$\begin{aligned}
\sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right)^2 & \leq \left\lfloor \frac{M}{\alpha} \right\rfloor \frac{(\alpha-1)^2}{M^2} + \left(M - \left\lfloor \frac{M}{\alpha} \right\rfloor \right) \frac{1}{M^2} \\
& \leq \frac{M}{\alpha} \frac{(\alpha-1)^2}{M^2} + \left(M - \frac{M}{\alpha} + 1 \right) \frac{1}{M^2} \\
& = \frac{\alpha-1}{M} + \frac{1}{M^2} \leq \frac{\alpha}{M}. \tag{6}
\end{aligned}$$

1016 We can transform (5) into

$$\mathbb{E}_{\xi_m^{k,h}} \left\| \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}) \right\|^2 \leq 4\alpha(\delta_1+1) \|\nabla f(x^{k,h})\|^2 + 4\alpha\delta_2 + \frac{2\alpha\sigma^2}{M}. \tag{7}$$

1017 Substituting (7) into (4), we have

$$\begin{aligned}
\mathbb{E}_{\xi_m^{k,h}} \|g^{k,h+1}\|^2 & \leq (1+c) \|g^{k,h}\|^2 + 4 \left(1 + \frac{1}{c} \right) (1-\theta)^2 \alpha(\delta_1+1) \|\nabla f(x^{k,h})\|^2 \\
& + 4 \left(1 + \frac{1}{c} \right) (1-\theta)^2 \alpha\delta_2 \\
& + 2 \left(1 + \frac{1}{c} \right) (1-\theta)^2 \frac{\alpha}{M} \sigma^2.
\end{aligned}$$

1018 Enrolling a recursion, we get

$$\begin{aligned}
\mathbb{E}_{\xi_m^{k,0}} \dots \mathbb{E}_{\xi_m^{k,h}} \|g^{k,h+1}\|^2 & \leq 4 \left(1 + \frac{1}{c} \right) (1-\theta)^2 \alpha(\delta_1+1) \sum_{i=0}^h (1+c)^{h-i} \|\nabla f(x^{k,i})\|^2 \\
& + 4 \left(1 + \frac{1}{c} \right) (1-\theta)^2 \alpha\delta_2 \sum_{i=0}^h (1+c)^{h-i} \\
& + 2 \left(1 + \frac{1}{c} \right) (1-\theta)^2 \frac{\alpha}{M} \sigma^2 \sum_{i=0}^h (1+c)^{h-i}. \tag{8}
\end{aligned}$$

1019 Now we use that $H^k \sim \text{Geom}(p)$:

$$\begin{aligned}
\mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|g^{k,H^k}\|^2 & = \sum_{j \geq 0} p(1-p)^j \mathbb{E}_{\xi_m^{k,0}} \dots \mathbb{E}_{\xi_m^{k,j-1}} \|g^{k,j}\|^2 \\
& \stackrel{(8)}{\leq} 4 \left(1 + \frac{1}{c} \right) (1-\theta)^2 \alpha(\delta_1+1).
\end{aligned}$$

$$\begin{aligned}
& \cdot \sum_{j \geq 0} p(1-p)^j \sum_{i=0}^{j-1} (1+c)^{j-i-1} \|\nabla f(x^{k,i})\|^2 \\
& + 2 \left(1 + \frac{1}{c}\right) (1-\theta)^2 \frac{\alpha}{M} (\sigma^2 + 2M\delta_2) \cdot \\
& \cdot \sum_{j \geq 0} p(1-p)^j \sum_{i=0}^{j-1} (1+c)^{j-i-1}.
\end{aligned} \tag{9}$$

1020 Let us choose $c = \frac{p}{2}$ and consider the following term individually:

$$\begin{aligned}
& \sum_{j \geq 0} p(1-p)^j \sum_{i=0}^{j-1} (1+c)^{j-i-1} \|\nabla f(x^{k,i})\|^2 = p \left[(1-p) \|\nabla f(x^{k,0})\|^2 \right. \\
& \quad \left. + (1-p)^2 \left\{ (1+c) \|\nabla f(x^{k,0})\|^2 + \|\nabla f(x^{k,1})\|^2 \right\} + \dots \right] \\
& = p(1-p) \left[(1-p)^0 (1+c)^0 + (1-p)(1+c) + \dots \right] \|\nabla f(x^{k,0})\|^2 \\
& \quad + p(1-p)^2 \left[(1-p)^0 (1+c)^0 + (1-p)(1+c) + \dots \right] \|\nabla f(x^{k,1})\|^2 + \dots \\
& \leq \sum_{l \geq 0} (1-p)^l \left(1 + \frac{p}{2}\right)^l \sum_{j \geq 0} p(1-p)^{j+1} \|\nabla f(x^{k,j})\|^2 \\
& \leq \frac{1}{1 - (1-p)(1 + \frac{p}{2})} \sum_{j \geq 0} p(1-p)^j \|\nabla f(x^{k,j})\|^2 = \frac{2}{p(p+1)} \mathbb{E}_{H^k} \|\nabla f(x^{k,H^k})\|^2 \\
& \leq \frac{2}{p} \mathbb{E}_{H^k} \|\nabla f(x^{k,H^k})\|^2.
\end{aligned} \tag{10}$$

1021 Additionally, we have

$$\begin{aligned}
\sum_{j \geq 0} p(1-p)^j \sum_{i=0}^{j-1} (1+c)^{j-i-1} & \leq p \sum_{j \geq 0} (1-p)^j j \left(1 + \frac{p}{2}\right)^j \leq p \sum_{j \geq 0} j \left(1 - \frac{p}{2}\right)^j \\
& = p \frac{1 - \frac{p}{2}}{\left(1 - \left(1 - \frac{p}{2}\right)\right)^2} \leq \frac{4}{p}.
\end{aligned} \tag{11}$$

1022 Combining this estimates with (9) we obtain the result of the lemma:

$$\begin{aligned}
\mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|g^{k,H^k}\|^2 & \leq \frac{24(1-\theta)^2 \alpha (\delta_1 + 1)}{p^2} \mathbb{E}_{H^k} \|\nabla f(x^{k,H^k})\|^2 + \frac{48(1-\theta)^2 \alpha \delta_2}{p^2} \\
& \quad + \frac{24(1-\theta)^2 \alpha \sigma^2}{Mp^2}.
\end{aligned}$$

1023 □

1024 D.1 Proof for non-convex case

1025 **Theorem 3 (Theorem 1).** Suppose Assumptions 1, 2(a), 3, 4 hold. Then for Algorithm 1 with
1026 $\theta \leq \frac{\gamma L p^2}{2}$ and $\gamma \leq \frac{p}{384 L \alpha (\delta_1 + 1)}$ it implies that

$$\begin{aligned}
\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(x^{k,0})\|^2 & \leq \frac{16(f(x^{0,0}) - f(x^*))}{\gamma K} + \frac{768 \gamma L \alpha \delta_2}{p} + \frac{384 \gamma^2 L^2 \alpha \delta_2}{p^3} \\
& \quad + \frac{400 \gamma L \alpha \sigma^2}{Mp} + \frac{192 \gamma^2 L^2 \alpha \sigma^2}{Mp^3}.
\end{aligned}$$

1027 *Proof.* We start with the definition of virtual sequence:

$$\tilde{x}^{k,h} = x^{k,h} - \gamma \sum_{m=1}^M g_m^{k,h} = x^{k,h} - \gamma g^{k,h}. \quad (12)$$

1028 It is followed by

$$\begin{aligned} \tilde{x}^{k,h+1} &= x^{k,h+1} - \gamma \sum_{m=1}^M g_m^{k,h+1} = x^{k,h} - \gamma \left[(1-\theta) \sum_{m=1}^M \tilde{\pi}_m^{k,h} \nabla f_m(x^{k,h}, \xi_m^{k,h}) + \theta g^{k-1, H^{k-1}} \right] \\ &\quad - \gamma \sum_{m=1}^M g_m^{k,h} - \gamma (1-\theta) \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \\ &= \tilde{x}^{k,h} - \gamma \left[(1-\theta) \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) + \theta g^{k-1, H^{k-1}} \right]. \end{aligned} \quad (13)$$

1029 Assumption 1 implies

$$\begin{aligned} f(\tilde{x}^{k,h+1}) &\leq f(\tilde{x}^{k,h}) + \langle \nabla f(\tilde{x}^{k,h}), \tilde{x}^{k,h+1} - \tilde{x}^{k,h} \rangle + \frac{L}{2} \|\tilde{x}^{k,h+1} - \tilde{x}^{k,h}\|^2 \\ &\stackrel{(13)}{\leq} f(\tilde{x}^{k,h}) - \gamma \theta \langle \nabla f(\tilde{x}^{k,h}), g^{k-1, H^{k-1}} \rangle \\ &\quad - \gamma (1-\theta) \left\langle \nabla f(\tilde{x}^{k,h}), \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ &\quad + \frac{\gamma^2 L (1-\theta)}{2} \left\| \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 + \frac{\gamma^2 L \theta}{2} \|g^{k-1, H^{k-1}}\|^2. \end{aligned}$$

1030 Taking expectation over $\xi_m^{k,h}$, we have

$$\begin{aligned} \mathbb{E}_{\xi_m^{k,h}} [f(\tilde{x}^{k,h+1})] &\leq \mathbb{E}_{\xi_m^{k,h}} [f(\tilde{x}^{k,h})] - \gamma \theta \mathbb{E}_{\xi_m^{k,h}} \langle \nabla f(\tilde{x}^{k,h}), g^{k-1, H^{k-1}} \rangle \\ &\quad - \gamma (1-\theta) \mathbb{E}_{\xi_m^{k,h}} \left\langle \nabla f(\tilde{x}^{k,h}), \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ &\quad + \frac{\gamma^2 L (1-\theta)}{2} \mathbb{E}_{\xi_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\ &\quad + \frac{\gamma^2 L \theta}{2} \mathbb{E}_{\xi_m^{k,h}} \|g^{k-1, H^{k-1}}\|^2. \end{aligned} \quad (14)$$

1031 Note that

$$\begin{aligned} \tilde{x}^{k,h} &\stackrel{(12)}{=} x^{k,h} - \gamma g^{k,h} \\ &\stackrel{\text{Line 12}}{=} x^{k,h} - \gamma \left(g^{k,h-1} + (1-\theta) \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h-1} \right) \nabla f(x^{k,h-1}, \xi_m^{k,h-1}) \right). \end{aligned}$$

1032 Thus, $\tilde{x}^{k,h}$ and $\xi_m^{k,h}$ are independent. Analogously, $g^{k-1, H^{k-1}}$ and $\xi_m^{k,h}$ are independent. In this way,
1033 (14) transforms into

$$\begin{aligned} \mathbb{E}_{\xi_m^{k,h}} [f(\tilde{x}^{k,h+1})] &\leq f(\tilde{x}^{k,h}) - \gamma \theta \langle \nabla f(\tilde{x}^{k,h}), g^{k-1, H^{k-1}} \rangle \\ &\quad - \gamma (1-\theta) \langle \nabla f(\tilde{x}^{k,h}), \nabla f(x^{k,h}) \rangle \\ &\quad + \frac{\gamma^2 L (1-\theta)}{2} \mathbb{E}_{\xi_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \end{aligned}$$

$$\begin{aligned}
& + \frac{\gamma^2 L \theta}{2} \|g^{k-1, H^{k-1}}\|^2 \\
\stackrel{\text{(CS)}}{\leq} & f(\tilde{x}^{k,h}) - \gamma \theta \langle \nabla f(\tilde{x}^{k,h}), g^{k-1, H^{k-1}} \rangle \\
& - \gamma(1-\theta) \langle \nabla f(\tilde{x}^{k,h}), \nabla f(x^{k,h}) \rangle \\
& + \gamma^2 L(1-\theta) \mathbb{E}_{\xi_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M (\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})) \right\|^2 \\
& + \gamma^2 L(1-\theta) \|\nabla f(x^{k,h})\|^2 \\
& + \frac{\gamma^2 L \theta}{2} \|g^{k-1, H^{k-1}}\|^2.
\end{aligned} \tag{15}$$

1034 Now we pay attention to the following term:

$$\begin{aligned}
& \mathbb{E}_{\xi_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M (\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})) \right\|^2 \\
& \stackrel{(i)}{=} \frac{1}{M^2} \sum_{m=1}^M \mathbb{E}_{\xi_m^{k,h}} \|\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})\|^2 \\
& + \frac{2}{M^2} \sum_{i \neq j} \left\langle \mathbb{E}_{\xi_i^{k,h}} [\nabla f_i(x^{k,h}, \xi_i^{k,h}) - \nabla f_i(x^{k,h})], \mathbb{E}_{\xi_j^{k,h}} [\nabla f_j(x^{k,h}, \xi_j^{k,h}) - \nabla f_j(x^{k,h})] \right\rangle \\
& \stackrel{\text{As. 4}}{\leq} \frac{1}{M} \sigma^2,
\end{aligned}$$

1035 where (i) is correct, since $\xi_i^{k,h}$ and $\xi_j^{k,h}$ are independent. Substituting this estimate into (15), we have

$$\begin{aligned}
\mathbb{E}_{\xi_m^{k,h}} [f(\tilde{x}^{k,h+1})] & \leq f(\tilde{x}^{k,h}) - \gamma \theta \langle \nabla f(\tilde{x}^{k,h}), g^{k-1, H^{k-1}} \rangle \\
& - \gamma(1-\theta) \langle \nabla f(\tilde{x}^{k,h}), \nabla f(x^{k,h}) \rangle \\
& + \gamma^2 L(1-\theta) \|\nabla f(x^{k,h})\|^2 + \frac{\gamma^2 L \theta}{2} \|g^{k-1, H^{k-1}}\|^2 \\
& + \frac{\gamma^2 L(1-\theta) \sigma^2}{M}.
\end{aligned} \tag{16}$$

1036 Let us estimate the scalar products separately.

$$\begin{aligned}
-\gamma(1-\theta) \langle \nabla f(\tilde{x}^{k,h}), \nabla f(x^{k,h}) \rangle & = -\frac{\gamma(1-\theta)}{2} \|\nabla f(\tilde{x}^{k,h})\|^2 - \frac{\gamma(1-\theta)}{2} \|\nabla f(x^{k,h})\|^2 \\
& + \frac{\gamma(1-\theta)}{2} \|\nabla f(\tilde{x}^{k,h}) - \nabla f(x^{k,h})\|^2 \\
& \stackrel{\text{As. 1}}{\leq} -\frac{\gamma(1-\theta)}{2} \|\nabla f(\tilde{x}^{k,h})\|^2 - \frac{\gamma(1-\theta)}{2} \|\nabla f(x^{k,h})\|^2 \\
& + \frac{\gamma L^2(1-\theta)}{2} \|\tilde{x}^{k,h} - x^{k,h}\|^2 \\
& \stackrel{(12)}{=} -\frac{\gamma(1-\theta)}{2} \|\nabla f(\tilde{x}^{k,h})\|^2 - \frac{\gamma(1-\theta)}{2} \|\nabla f(x^{k,h})\|^2 \\
& + \frac{\gamma^3 L^2(1-\theta)}{2} \|g^{k,h}\|^2, \\
-\gamma \theta \langle \nabla f(\tilde{x}^{k,h}), g^{k-1, H^{k-1}} \rangle & \stackrel{\text{(Fen)}}{\leq} \frac{\gamma \theta}{2} \|\nabla f(\tilde{x}^{k,h})\|^2 + \frac{\gamma \theta}{2} \|g^{k-1, H^{k-1}}\|^2.
\end{aligned}$$

1037 Combining it with (16), we have

$$\begin{aligned}\mathbb{E}_{\xi_m^{k,h}} [f(\tilde{x}^{k,h+1})] &\leq f(\tilde{x}^{k,h}) - \frac{\gamma(1-\theta)}{2} (1-2\gamma L) \|\nabla f(x^{k,h})\|^2 - \frac{\gamma(1-2\theta)}{2} \|\nabla f(\tilde{x}^{k,h})\|^2 \\ &\quad + \frac{\gamma^3 L^2 (1-\theta)}{2} \|g^{k,h}\|^2 + \frac{\gamma\theta(\gamma L+1)}{2} \|g^{k-1,H^{k-1}}\|^2 + \frac{\gamma^2 L(1-\theta)\sigma^2}{M}.\end{aligned}$$

1038 Now we put $h = H^k - 1$ and take additional expectations.

$$\begin{aligned}&\mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} [f(\tilde{x}^{k,H^k})] \\ &\leq \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} [f(\tilde{x}^{k,H^k-1})] \\ &\quad - \frac{\gamma(1-\theta)}{2} (1-2\gamma L) \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|\nabla f(x^{k,H^k-1})\|^2 \\ &\quad - \frac{\gamma(1-2\theta)}{2} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|\nabla f(\tilde{x}^{k,H^k-1})\|^2 \\ &\quad + \frac{\gamma^3 L^2 (1-\theta)}{2} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|g^{k,H^k-1}\|^2 \\ &\quad + \frac{\gamma\theta(\gamma L+1)}{2} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^k-1-1}} \|g^{k-1,H^k-1}\|^2 \\ &\quad + \frac{\gamma^2 L(1-\theta)\sigma^2}{M}.\end{aligned}$$

1039 We take expectation with respect to H^{k-1} and H^k , and apply Lemma 2:

$$\begin{aligned}&\mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} [f(\tilde{x}^{k,H^k})] \\ &\leq (1-p) \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} [f(\tilde{x}^{k,H^k})] \\ &\quad + p \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^k-1-1}} [f(\tilde{x}^{k,0})] \\ &\quad - \frac{\gamma(1-\theta)p}{2} (1-2\gamma L) \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^k-1-1}} \|\nabla f(x^{k,0})\|^2 \\ &\quad - \frac{\gamma(1-\theta)(1-p)}{2} (1-2\gamma L) \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|\nabla f(x^{k,H^k})\|^2 \\ &\quad - \frac{\gamma(1-2\theta)p}{2} \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^k-1-1}} \|\nabla f(\tilde{x}^{k,0})\|^2 \\ &\quad - \frac{\gamma(1-2\theta)(1-p)}{2} \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|\nabla f(\tilde{x}^{k,H^k})\|^2 \\ &\quad + \frac{\gamma^3 L^2 (1-\theta)p}{2} \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^k-1-1}} \underbrace{\|g^{k,0}\|^2}_{=0} \\ &\quad + \frac{\gamma^3 L^2 (1-\theta)(1-p)}{2} \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|g^{k,H^k}\|^2 \\ &\quad + \frac{\gamma\theta(\gamma L+1)}{2} \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^k-1-1}} \|g^{k-1,H^k-1}\|^2 \\ &\quad + \frac{\gamma^2 L(1-\theta)\sigma^2}{M}.\end{aligned}$$

1040 Next, we put $\gamma \leq \frac{1}{4L}$ and $\theta \leq \frac{1}{2}$. Moreover, we use that H^k and $\{\xi_m^{k-1,h}\}_{h=0}^{H^{k-1}-1}$ are independent
1041 stochastic values.

$$\begin{aligned}&\mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} [f(\tilde{x}^{k,H^k})] \\ &\leq (1-p) \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} [f(\tilde{x}^{k,H^k})]\end{aligned}$$

$$\begin{aligned}
& + p\mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} [f(\tilde{x}^{k,0})] \\
& - \frac{\gamma(1-\theta)p}{4}\mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \|\nabla f(x^{k,0})\|^2 \\
& - \frac{\gamma(1-\theta)(1-p)}{4}\mathbb{E}_{H^{k-1}}\mathbb{E}_{H^k}\mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|\nabla f(x^{k,H^k})\|^2 \\
& + \frac{\gamma^3 L^2(1-\theta)(1-p)}{2}\mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}}\mathbb{E}_{H^k}\mathbb{E}_{\xi_m^{k,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|g^{k,H^k}\|^2 \\
& + \gamma\theta\mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \|g^{k-1,H^{k-1}}\|^2 \\
& + \frac{\gamma^2 L(1-\theta)\sigma^2}{M}.
\end{aligned} \tag{17}$$

1042 We use Lemma 1 to estimate $\|g^{k,H^k}\|^2$ and $\|g^{k-1,H^{k-1}}\|^2$. We obtain

$$\begin{aligned}
\mathbb{E}_{H^k}\mathbb{E}_{\xi_m^{k,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|g^{k,H^k}\|^2 & \leq \frac{24(1-\theta)^2\alpha(\delta_1+1)}{p^2}\mathbb{E}_{H^k} \|\nabla f(x^{k,H^k})\|^2 + \frac{48(1-\theta)^2\alpha\delta_2}{p^2} \\
& + \frac{24(1-\theta)^2\alpha\sigma^2}{Mp^2}.
\end{aligned} \tag{18}$$

1043 As for $\|g^{k-1,H^{k-1}}\|^2$, we have

$$\begin{aligned}
& \mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \|g^{k-1,H^{k-1}}\|^2 \\
& \leq \frac{24(1-\theta)^2\alpha(\delta_1+1)}{p^2}\mathbb{E}_{H^{k-1}} \|\nabla f(x^{k-1,H^{k-1}})\|^2 + \frac{48(1-\theta)^2\alpha\delta_2}{p^2} + \frac{24(1-\theta)^2\alpha\sigma^2}{Mp^2} \\
& \stackrel{\text{(CS)}}{\leq} \frac{48(1-\theta)\alpha(\delta_1+1)}{p^2}\mathbb{E}_{H^{k-1}} \|\nabla f(x^{k-1,H^{k-1}}) - \nabla f(\tilde{x}^{k-1,H^{k-1}})\|^2 \\
& \quad + \frac{48(1-\theta)^2\alpha(\delta_1+1)}{p^2}\mathbb{E}_{H^{k-1}} \|\nabla f(\tilde{x}^{k-1,H^{k-1}})\|^2 + \frac{48(1-\theta)^2\alpha\delta_2}{p^2} + \frac{24(1-\theta)^2\alpha\sigma^2}{Mp^2} \\
& \stackrel{\text{As. 1}}{\leq} \frac{48L^2(1-\theta)^2\alpha(\delta_1+1)}{p^2}\mathbb{E}_{H^{k-1}} \|x^{k-1,H^{k-1}} - \tilde{x}^{k-1,H^{k-1}}\|^2 \\
& \quad + \frac{48(1-\theta)^2\alpha(\delta_1+1)}{p^2} \|\nabla f(x^{k,0})\|^2 + \frac{48(1-\theta)^2\alpha\delta_2}{p^2} + \frac{24(1-\theta)^2\alpha\sigma^2}{Mp^2} \\
& \stackrel{\text{(12)}}{=} \frac{48\gamma^2 L^2(1-\theta)^2\alpha(\delta_1+1)}{p^2}\mathbb{E}_{H^{k-1}} \|g^{k-1,H^{k-1}}\|^2 \\
& \quad + \frac{48(1-\theta)^2\alpha(\delta_1+1)}{p^2} \|\nabla f(x^{k,0})\|^2 + \frac{48(1-\theta)^2\alpha\delta_2}{p^2} + \frac{24(1-\theta)^2\alpha\sigma^2}{Mp^2}.
\end{aligned}$$

1044 We choose $\gamma \leq \frac{p}{96L\sqrt{\alpha\sqrt{\delta_1+1}}}$. Moreover, we take additional expectations and again use that H^{k-1}

1045 and $\{\xi_m^{k-1,h}\}_{h=0}^{H^{k-1}-1}$ are independent stochastic values:

$$\begin{aligned}
& \mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \|g^{k-1,H^{k-1}}\|^2 \\
& \leq \frac{96(1-\theta)^2\alpha(\delta_1+1)}{p^2}\mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \|\nabla f(x^{k,0})\|^2 + \frac{96(1-\theta)^2\alpha\delta_2}{p^2} \\
& \quad + \frac{48(1-\theta)^2\alpha\sigma^2}{Mp^2}.
\end{aligned} \tag{19}$$

1046 Now we substitute (18) and (19) into (17):

$$\begin{aligned}
& p\mathbb{E}_{H^{k-1}}\mathbb{E}_{H^k}\mathbb{E}_{\xi_m^{k-1,0}}\dots\mathbb{E}_{\xi_m^{k,H^k-1}}\left[f(\tilde{x}^{k,H^k})\right] \\
& \leq p\mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}}\dots\mathbb{E}_{\xi_m^{k-1,H^k-1-1}}\left[f(\tilde{x}^{k,0})\right] \\
& \quad - \frac{\gamma(1-\theta)p}{4}\mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}}\dots\mathbb{E}_{\xi_m^{k-1,H^k-1-1}}\left\|\nabla f(x^{k,0})\right\|^2 \\
& \quad - \frac{\gamma(1-\theta)(1-p)}{4}\mathbb{E}_{H^{k-1}}\mathbb{E}_{H^k}\mathbb{E}_{\xi_m^{k-1,0}}\dots\mathbb{E}_{\xi_m^{k,H^k-1}}\left\|\nabla f(x^{k,H^k})\right\|^2 \\
& \quad + \frac{12\gamma^3L^2(1-\theta)^3(1-p)\alpha(\delta_1+1)}{p^2}\mathbb{E}_{H^{k-1}}\mathbb{E}_{\xi_m^{k-1,0}}\dots\mathbb{E}_{\xi_m^{k-1,H^k-1-1}}\mathbb{E}_{H^k}\left\|\nabla f(x^{k,H^k})\right\|^2 \\
& \quad + \frac{24\gamma^3L^2(1-\theta)^3(1-p)\alpha\delta_2}{p^2} + \frac{12\gamma^3L^2(1-\theta)^3(1-p)\alpha\sigma^2}{Mp^2} \\
& \quad + \frac{96\gamma\theta(1-\theta)^2\alpha(\delta_1+1)}{p^2}\mathbb{E}_{\xi_m^{k-1,0}}\dots\mathbb{E}_{\xi_m^{k-1,H^k-1-1}}\left\|\nabla f(x^{k,0})\right\|^2 \\
& \quad + \frac{96\gamma\theta(1-\theta)^2\alpha\delta_2}{p^2} + \frac{48\gamma\theta(1-\theta)^2\alpha\sigma^2}{Mp^2} + \frac{\gamma^2L(1-\theta)\sigma^2}{M}.
\end{aligned}$$

1047 We take the full expectation, then use a law of expectation and rearrange terms:

$$\begin{aligned}
p\mathbb{E}\left[f(\tilde{x}^{k,H^k})\right] & \leq p\mathbb{E}\left[f(\tilde{x}^{k,0})\right] \\
& \quad - \frac{\gamma(1-\theta)(1-p)}{4}\left(1 - \frac{48\gamma^2L^2(1-\theta)^2\alpha(\delta_1+1)}{p^2}\right)\mathbb{E}\left\|\nabla f(x^{k,H^k})\right\|^2 \\
& \quad - \frac{\gamma(1-\theta)p}{4}\left(1 - \frac{384\theta(1-\theta)\alpha(\delta_1+1)}{p^3}\right)\mathbb{E}\left\|\nabla f(x^{k,0})\right\|^2 \\
& \quad + \frac{24\gamma^3L^2(1-\theta)^3(1-p)\alpha\delta_2}{p^2} + \frac{96\gamma\theta(1-\theta)^2\alpha\delta_2}{p^2} \\
& \quad + \frac{\gamma^2L(1-\theta)\sigma^2}{M} + \frac{12\gamma^3L^2(1-\theta)^3(1-p)\alpha\sigma^2}{Mp^2} + \frac{48\gamma\theta(1-\theta)^2\alpha\sigma^2}{Mp^2}.
\end{aligned}$$

1048 We choose $\theta \leq \frac{\gamma L p^2}{2}$ and $\gamma \leq \frac{p}{384L\alpha(\delta_1+1)}$. Note that all previous transitions hold even with larger
1049 choice of θ and γ , consequently this choice is correct. In that way, we obtain

$$\begin{aligned}
\frac{\gamma(1-\theta)p}{8}\mathbb{E}\left\|\nabla f(x^{k,0})\right\|^2 & \leq p\mathbb{E}\left[f(\tilde{x}^{k,0}) - f(\tilde{x}^{k,H^k})\right] \\
& \quad + \frac{24\gamma^3L^2\alpha\delta_2}{p^2} + 48\gamma^2L\alpha\delta_2 \\
& \quad + \frac{\gamma^2L\sigma^2}{M} + \frac{12\gamma^3L^2\alpha\sigma^2}{Mp^2} + \frac{24\gamma^2L\alpha\sigma^2}{M}.
\end{aligned}$$

1050 Note that $\tilde{x}^{k,H^k} = x^{k,H^k} - \gamma g^{k,H^k} = x^{k+1,0}$ and $\tilde{x}^{k,0} = x^{k,0}$. Thus,

$$\begin{aligned}
\frac{\gamma(1-\theta)}{8}\mathbb{E}\left\|\nabla f(x^{k,0})\right\|^2 & \leq \mathbb{E}\left[f(x^{k,0}) - f(x^{k+1,0})\right] + \frac{48\gamma L\alpha\delta_2}{p} + \frac{24\gamma^2L^2\alpha\delta_2}{p^3} \\
& \quad + \frac{25\gamma^2L\alpha\sigma^2}{Mp} + \frac{12\gamma^3L^2\alpha\sigma^2}{Mp^3}.
\end{aligned}$$

1051 Averaging over all epochs, we obtain the result of the theorem:

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(x^{k,0})\right\|^2 \leq \frac{8(f(x^{0,0}) - \mathbb{E}[f(x^{K,0})])}{\gamma(1-\theta)K} + \frac{384\gamma L\alpha\delta_2}{p(1-\theta)} + \frac{192\gamma^2L^2\alpha\delta_2}{p^3(1-\theta)}$$

$$\begin{aligned}
& + \frac{200\gamma L\alpha\sigma^2}{Mp(1-\theta)} + \frac{96\gamma^2 L^2\alpha\sigma^2}{Mp^3(1-\theta)} \\
\leq & \frac{16(f(x^{0,0}) - f(x^*))}{\gamma K} + \frac{768\gamma L\alpha\delta_2}{p} + \frac{384\gamma^2 L^2\alpha\delta_2}{p^3} \\
& + \frac{400\gamma L\alpha\sigma^2}{Mp} + \frac{192\gamma^2 L^2\alpha\sigma^2}{Mp^3}.
\end{aligned}$$

1052

□

1053 **Corollary 5 (Corollary 1).** *Under conditions of Theorem 1 Algorithm 1 with fixed rules $\widehat{\mathcal{R}}^k \equiv$*
1054 *$\widehat{\mathcal{R}}^{k,h} \equiv \mathcal{R}$ needs*

$$\begin{aligned}
& \mathcal{O}\left(\frac{M}{C}\left(\frac{\Delta L\alpha\delta_1}{\varepsilon^2} + \frac{\Delta L\alpha\delta_2}{\varepsilon^4} + \frac{\Delta L\alpha\sigma^2}{M\varepsilon^4}\right)\right) \text{ epochs and} \\
& \mathcal{O}\left(M\frac{M}{C}\left(\frac{\Delta L\alpha\delta_1}{\varepsilon^2} + \frac{\Delta L\alpha\delta_2}{\varepsilon^4} + \frac{\Delta L\alpha\sigma^2}{M\varepsilon^4}\right)\right) \text{ number of devices communications}
\end{aligned}$$

1055 *to reach ε -accuracy, where $\varepsilon^2 = \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(x^{k,0})\|^2$, $\Delta = f(x^{0,0}) - f(x^*)$ and C is the number*
1056 *of devices participating in each epoch.*

1057 *Proof.* Using the result of Theorem 1, we choose

$$\gamma \leq \min \left\{ \frac{p}{384L\alpha(\delta_1 + 1)}, \frac{\sqrt{(f(x^{0,0}) - f(x^*))p}}{4\sqrt{3}L\alpha\delta_2 K}, \frac{\sqrt[3]{(f(x^{0,0}) - f(x^*))p}}{2\sqrt[3]{3}L^2\alpha\delta_2 K}, \right. \\
\left. \frac{\sqrt{(f(x^{0,0}) - f(x^*))Mp}}{5\sqrt{L\alpha\sigma^2 K}}, \frac{\sqrt[3]{(f(x^{0,0}) - f(x^*))Mp}}{\sqrt[3]{12}L^2\alpha\sigma^2 K} \right\}.$$

1058 Thus, we need

$$\begin{aligned}
& \mathcal{O}\left(\frac{(f(x^{0,0}) - f(x^*))L\alpha\delta_1}{p\varepsilon^2} + \frac{(f(x^{0,0}) - f(x^*))L\alpha\delta_2}{p\varepsilon^4} + \frac{(f(x^{0,0}) - f(x^*))L\alpha\sigma^2}{Mp\varepsilon^4}\right. \\
& \left. + \frac{(f(x^{0,0}) - f(x^*))L\sqrt{\alpha\delta_2}}{p^{\frac{3}{2}}\varepsilon^3} + \frac{(f(x^{0,0}) - f(x^*))L\sqrt{\alpha\sigma}}{\sqrt{Mp}p^{\frac{3}{2}}\varepsilon^3}\right)
\end{aligned}$$

1059 epochs to reach ε -accuracy, where $\varepsilon^2 = \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(x^{k,0})\|^2$. Since the last two

1060 terms in the estimate in a magnitude smaller, than the second and third accordingly,
1061 we can ignore them. The length of the epoch $H \in \text{Geom}(p)$, Algorithm 1 requires

1062 $\mathcal{O}\left(\frac{(f(x^{0,0}) - f(x^*))L\alpha\delta_1}{p^2\varepsilon^2} + \frac{(f(x^{0,0}) - f(x^*))L\alpha\delta_2}{p^2\varepsilon^4} + \frac{(f(x^{0,0}) - f(x^*))L\alpha\sigma^2}{Mp^2\varepsilon^4}\right)$ communication rounds.

1063 Next we mention that at each communication round we communicate with C devices, thus, number of

1064 communications is $\mathcal{O}\left(C\frac{(f(x^{0,0}) - f(x^*))L\alpha\delta_1}{p^2\varepsilon^2} + C\frac{(f(x^{0,0}) - f(x^*))L\alpha\delta_2}{p^2\varepsilon^4} + C\frac{(f(x^{0,0}) - f(x^*))L\alpha\sigma^2}{Mp^2\varepsilon^4}\right)$.

1065 Taking $p = \frac{C}{M}$, we have the result of the corollary. The choice of p is motivated by the fact
1066 that we perform $\frac{1}{p}C + M$ communications per-epoch, and established p is the minimal, which
1067 delivers $\mathcal{O}(M)$ communications at each epoch. This is also the reason for the additional factor M in
1068 the estimate on communications. □

1069 **Corollary 6.** *Under conditions of Theorem 1 Algorithm 1 needs*

$$\begin{aligned}
& \mathcal{O}\left(\frac{M}{\min_{k,h} C^{k,h}}\left(\frac{\Delta L\alpha\delta_1}{\varepsilon^2} + \frac{\Delta L\alpha\delta_2}{\varepsilon^4} + \frac{\Delta L\alpha\sigma^2}{M\varepsilon^4}\right)\right) \text{ epochs and} \\
& \mathcal{O}\left(M\left(\frac{M}{\min_{k,h} C^{k,h}}\right)^2\left(\frac{\Delta L\alpha\delta_1}{\varepsilon^2} + \frac{\Delta L\alpha\delta_2}{\varepsilon^4} + \frac{\Delta L\alpha\sigma^2}{M\varepsilon^4}\right)\right) \text{ number of devices communications}
\end{aligned}$$

1070 to reach ε -accuracy, where $\varepsilon^2 = \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(x^{k,0})\|^2$, $\Delta = f(x^{0,0}) - f(x^*)$ and $C^{k,h}$ is the
 1071 number of devices participating in k -th iteration in h -th epoch.

1072 *Proof.* Using the result of Theorem 1, we choose

$$\gamma \leq \min \left\{ \frac{p}{384L\alpha(\delta_1 + 1)}, \frac{\sqrt{(f(x^{0,0}) - f(x^*))p}}{4\sqrt{3L\alpha\delta_2 K}}, \frac{\sqrt[3]{(f(x^{0,0}) - f(x^*))p}}{2\sqrt[3]{3L^2\alpha\delta_2 K}}, \right. \\ \left. \frac{\sqrt{(f(x^{0,0}) - f(x^*))Mp}}{5\sqrt{L\alpha\sigma^2 K}}, \frac{\sqrt[3]{(f(x^{0,0}) - f(x^*))Mp}}{\sqrt[3]{12L^2\alpha\sigma^2 K}} \right\}.$$

1073 Thus, we need

$$\mathcal{O} \left(\frac{(f(x^{0,0}) - f(x^*))L\alpha\delta_1}{p\varepsilon^2} + \frac{(f(x^{0,0}) - f(x^*))L\alpha\delta_2}{p\varepsilon^4} + \frac{(f(x^{0,0}) - f(x^*))L\alpha\sigma^2}{Mp\varepsilon^4} \right. \\ \left. + \frac{(f(x^{0,0}) - f(x^*))L\sqrt{\alpha\delta_2}}{p^{\frac{3}{2}}\varepsilon^3} + \frac{(f(x^{0,0}) - f(x^*))L\sqrt{\alpha\sigma}}{\sqrt{Mp}p^{\frac{3}{2}}\varepsilon^3} \right)$$

1074 epochs to reach ε -accuracy, where $\varepsilon^2 = \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(x^{k,0})\|^2$. Since the last
 1075 two terms in the estimate in a magnitude smaller, than the second and third ac-
 1076 cordingly, we can ignore them. The length of the epoch $H \in \text{Geom}(p)$, Algo-
 1077 rithm 1 requires $\mathcal{O} \left(\frac{(f(x^{0,0}) - f(x^*))L\alpha\delta_1}{p^2\varepsilon^2} + \frac{(f(x^{0,0}) - f(x^*))L\alpha\delta_2}{p^2\varepsilon^4} + \frac{(f(x^{0,0}) - f(x^*))L\alpha\sigma^2}{Mp^2\varepsilon^4} \right)$
 1078 communication rounds. Next we mention that at each communication round
 1079 we communicate with $C^{k,h}$ devices, thus, number of communications is
 1080 $\mathcal{O} \left(\max_{k,h} C^{k,h} \left(\frac{(f(x^{0,0}) - f(x^*))L\alpha\delta_1}{p^2\varepsilon^2} + \frac{(f(x^{0,0}) - f(x^*))L\alpha\delta_2}{p^2\varepsilon^4} + \frac{(f(x^{0,0}) - f(x^*))L\alpha\sigma^2}{Mp^2\varepsilon^4} \right) \right)$. Taking,
 1081 $p = \frac{\min_{k,h} C^{k,h}}{M}$, we have the result of the corollary. The choice of p is motivated by the fact that we
 1082 perform $\frac{1}{p} \max_{k,h} C^{k,h} + M$ communications per-epoch, and established p is the minimal, which
 1083 delivers $\mathcal{O} \left(M \frac{M}{\min_{k,h} C^{k,h}} \right)$ communications at each epoch while guarantee the epoch is executed
 1084 (if we take $p = \frac{\max_{k,h} C^{k,h}}{M}$, we can meet $p = 1$). This is also the reason for the additional factor
 1085 $M \frac{M}{\min_{k,h} C^{k,h}}$ in the estimate on communications. \square

1086 **Remark 1.** Considering fixed rules $\widehat{\mathcal{R}} \equiv \widetilde{\mathcal{R}} \equiv \mathcal{R}$, we have $\mathcal{O} \left(M \frac{M}{C} \left(\frac{\Delta L \delta_1}{\varepsilon^2} + \frac{\Delta L \delta_2}{\varepsilon^4} + \frac{\Delta L \sigma^2}{M \varepsilon^4} \right) \right)$
 1087 and $\mathcal{O} \left(M^2 \frac{M}{C} \left(\frac{\Delta L \delta_1}{\varepsilon^2} + \frac{\Delta L \delta_2}{\varepsilon^4} + \frac{\Delta L \sigma^2}{M \varepsilon^4} \right) \right)$ number of devices communications with regularizing
 1088 parameter $\alpha = 1$ and $\alpha = M$ respectively. Considering various rules, best case with regularizing
 1089 coefficient $\alpha = 1$ gives us $\mathcal{O} \left(M \left(\frac{M}{\min_{k,h} C^{k,h}} \right)^2 \left(\frac{\Delta L \delta_1}{\varepsilon^2} + \frac{\Delta L \delta_2}{\varepsilon^4} + \frac{\Delta L \sigma^2}{M \varepsilon^4} \right) \right)$ and worst case $\alpha = M$
 1090 gives us $\mathcal{O} \left(M^2 \left(\frac{M}{\max_{k,h} C^{k,h}} \right)^2 \left(\frac{\Delta L \delta_1}{\varepsilon^2} + \frac{\Delta L \delta_2}{\varepsilon^4} + \frac{\Delta L \sigma^2}{M \varepsilon^4} \right) \right)$ number of devices communications.

1091 D.2 Proof for strongly-convex case

1092 **Theorem 4 (Theorem 2).** Suppose Assumptions 1, 2(b), 3, 4 hold. Then for Algorithm 1 with
 1093 $\theta \leq \frac{p\gamma\mu}{4}$ and $\gamma \leq \frac{p^2}{96L\alpha(\delta_1+1)}$ it implies that

$$\mathbb{E} \|x^{K,0} - x^*\|^2 \leq \left(1 - \frac{\gamma\mu}{8}\right)^K \|x^{0,0} - x^*\|^2 + \frac{8\gamma\alpha}{\mu p^3} \left(144\delta_2 + \frac{74\sigma^2}{M}\right).$$

1094 *Proof.* We start with the definition of virtual sequence:

$$\tilde{x}^{k,h} = x^{k,h} - \gamma \sum_{m=1}^M g_m^{k,h}. \quad (20)$$

1095 It is followed by

$$\begin{aligned} \tilde{x}^{k,h+1} &= x^{k,h+1} - \gamma \sum_{m=1}^M g_m^{k,h+1} \\ &= x^{k,h} - \gamma \left[(1-\theta) \sum_{m=1}^M \tilde{\pi}_m^{k,h} \nabla f_m(x^{k,h}, \xi_m^{k,h}) + \theta g^{k-1,H^{k-1}} \right] \\ &\quad - \gamma \sum_{m=1}^M g_m^{k,h} - \gamma (1-\theta) \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \\ &= \tilde{x}^{k,h} - \gamma \left[(1-\theta) \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) + \theta g^{k-1,H^{k-1}} \right]. \end{aligned} \quad (21)$$

1096 We use this to write a descent:

$$\begin{aligned} \|\tilde{x}^{k,h+1} - x^*\|^2 &= \|\tilde{x}^{k,h} - x^*\|^2 + 2 \langle \tilde{x}^{k,h} - x^*, \tilde{x}^{k,h+1} - \tilde{x}^{k,h} \rangle + \|\tilde{x}^{k,h+1} - \tilde{x}^{k,h}\|^2 \\ &\stackrel{(21)}{=} \|\tilde{x}^{k,h} - x^*\|^2 - 2\gamma\theta \langle \tilde{x}^{k,h} - x^*, g^{k-1,H^{k-1}} \rangle \\ &\quad - 2\gamma(1-\theta) \left\langle \tilde{x}^{k,h} - x^*, \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ &\quad + \gamma^2 \left\| \theta g^{k-1,H^{k-1}} + (1-\theta) \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\ &\stackrel{(\text{Jen})}{\leq} \|\tilde{x}^{k,h} - x^*\|^2 - 2\gamma\theta \langle \tilde{x}^{k,h} - x^*, g^{k-1,H^{k-1}} \rangle \\ &\quad - 2\gamma(1-\theta) \left\langle \tilde{x}^{k,h} - x^{k,h}, \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ &\quad - 2\gamma(1-\theta) \left\langle x^{k,h} - x^*, \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ &\quad + \gamma^2 \theta \|g^{k-1,H^{k-1}}\|^2 + \gamma^2 (1-\theta) \left\| \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2. \end{aligned}$$

1097 Taking the expectation over $\xi_m^{k,h}$, we have

$$\begin{aligned} \mathbb{E}_{\xi_m^{k,h}} \|\tilde{x}^{k,h+1} - x^*\|^2 &\leq \mathbb{E}_{\xi_m^{k,h}} \|\tilde{x}^{k,h} - x^*\|^2 - 2\gamma\theta \mathbb{E}_{\xi_m^{k,h}} \langle \tilde{x}^{k,h} - x^*, g^{k-1,H^{k-1}} \rangle \\ &\quad - 2\gamma(1-\theta) \mathbb{E}_{\xi_m^{k,h}} \left\langle \tilde{x}^{k,h} - x^{k,h}, \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ &\quad - 2\gamma(1-\theta) \mathbb{E}_{\xi_m^{k,h}} \left\langle x^{k,h} - x^*, \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ &\quad + \gamma^2 \theta \mathbb{E}_{\xi_m^{k,h}} \|g^{k-1,H^{k-1}}\|^2 \\ &\quad + \gamma^2 (1-\theta) \mathbb{E}_{\xi_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2. \end{aligned} \quad (22)$$

1098 Mention that

$$\begin{aligned}\tilde{x}^{k,h} &\stackrel{(20)}{=} x^{k,h} - \gamma g^{k,h} \\ &\stackrel{\text{Line 12}}{=} x^{k,h} - \gamma \left(g^{k,h-1} + (1-\theta) \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h-1} \right) \nabla f(x^{k,h-1}, \xi_m^{k,h-1}) \right),\end{aligned}$$

1099 Thus, $\tilde{x}^{k,h}$ and $\xi_m^{k,h}$ are independent. Analogously, $g^{k-1,H^{k-1}}$ and $\xi_m^{k,h}$ are independent. In this way,
1100 (22) transforms into

$$\begin{aligned}\mathbb{E}_{\xi_m^{k,h}} \|\tilde{x}^{k,h+1} - x^*\|^2 &\leq \|\tilde{x}^{k,h} - x^*\|^2 - 2\gamma\theta \langle \tilde{x}^{k,h} - x^*, g^{k-1,H^{k-1}} \rangle \\ &\quad - 2\gamma(1-\theta) \langle \tilde{x}^{k,h} - x^{k,h}, \nabla f(x^{k,h}) \rangle \\ &\quad - 2\gamma(1-\theta) \langle x^{k,h} - x^*, \nabla f(x^{k,h}) \rangle \\ &\quad + \gamma^2\theta \|g^{k-1,H^{k-1}}\|^2 \\ &\quad + \gamma^2(1-\theta) \mathbb{E}_{\xi_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\ &\stackrel{(\text{CS})}{\leq} \|\tilde{x}^{k,h} - x^*\|^2 - 2\gamma\theta \langle \tilde{x}^{k,h} - x^*, g^{k-1,H^{k-1}} \rangle \\ &\quad - 2\gamma(1-\theta) \langle \tilde{x}^{k,h} - x^{k,h}, \nabla f(x^{k,h}) \rangle \\ &\quad - 2\gamma(1-\theta) \langle x^{k,h} - x^*, \nabla f(x^{k,h}) \rangle \\ &\quad + 2\gamma^2(1-\theta) \mathbb{E}_{\xi_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M (\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f(x^{k,h})) \right\|^2 \\ &\quad + 2\gamma^2(1-\theta) \|\nabla f(x^{k,h})\|^2 + \gamma^2\theta \|g^{k-1,H^{k-1}}\|^2.\end{aligned}\tag{23}$$

1101 Now we pay attention to the following term:

$$\begin{aligned}&\mathbb{E}_{\xi_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M (\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})) \right\|^2 \\ &\stackrel{(i)}{=} \frac{1}{M^2} \sum_{m=1}^M \mathbb{E}_{\xi_m^{k,h}} \|\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})\|^2 \\ &\quad + \frac{2}{M^2} \sum_{i \neq j} \left\langle \mathbb{E}_{\xi_i^{k,h}} [\nabla f_i(x^{k,h}, \xi_i^{k,h}) - \nabla f_i(x^{k,h})], \mathbb{E}_{\xi_j^{k,h}} [\nabla f_j(x^{k,h}, \xi_j^{k,h}) - \nabla f_j(x^{k,h})] \right\rangle \\ &\stackrel{\text{As. 4}}{\leq} \frac{1}{M} \sigma^2,\end{aligned}$$

1102 where (i) is correct, since $\xi_i^{k,h}$ and $\xi_j^{k,h}$ are independent. Substituting this estimate into (23), we have

$$\begin{aligned}\mathbb{E}_{\xi_m^{k,h}} \|\tilde{x}^{k,h+1} - x^*\|^2 &\leq \|\tilde{x}^{k,h} - x^*\|^2 - 2\gamma\theta \langle \tilde{x}^{k,h} - x^*, g^{k-1,H^{k-1}} \rangle \\ &\quad - 2\gamma(1-\theta) \langle \tilde{x}^{k,h} - x^{k,h}, \nabla f(x^{k,h}) \rangle \\ &\quad - 2\gamma(1-\theta) \langle x^{k,h} - x^*, \nabla f(x^{k,h}) \rangle \\ &\quad + 2\gamma^2(1-\theta) \|\nabla f(x^{k,h})\|^2 + \gamma^2\theta \|g^{k-1,H^{k-1}}\|^2 \\ &\quad + \frac{2\gamma^2(1-\theta)\sigma^2}{M}.\end{aligned}\tag{24}$$

1103 Let us estimate scalar products separately.

$$\begin{aligned}
-2\gamma\theta \langle \tilde{x}^{k,h} - x^*, g^{k-1, H^{k-1}} \rangle &\stackrel{\text{(Fen)}}{\leq} \theta \|\tilde{x}^{k,h} - x^*\|^2 + \gamma^2\theta \|g^{k-1, H^{k-1}}\|^2, \\
-2\gamma(1-\theta) \langle \tilde{x}^{k,h} - x^{k,h}, \nabla f(x^{k,h}) \rangle &\stackrel{\text{(Fen)}}{\leq} (1-\theta) \|\tilde{x}^{k,h} - x^{k,h}\|^2 \\
&\quad + \gamma^2(1-\theta) \|\nabla f(x^{k,h})\|^2 \\
&\stackrel{(20)}{=} \gamma^2(1-\theta) \|g^{k,h}\|^2 + \gamma^2(1-\theta) \|\nabla f(x^{k,h})\|^2, \\
-2\gamma(1-\theta) \langle x^{k,h} - x^*, \nabla f(x^{k,h}) \rangle &\stackrel{\text{As. 2(b)}}{\leq} -\gamma\mu(1-\theta) \|x^{k,h} - x^*\|^2 \\
&\quad - 2\gamma(1-\theta) [f(x^{k,h}) - f(x^*)] \\
&\stackrel{\text{(CS)}}{\leq} -\frac{\gamma\mu(1-\theta)}{2} \|\tilde{x}^{k,h} - x^*\|^2 \\
&\quad + \gamma\mu(1-\theta) \|x^{k,h} - \tilde{x}^{k,h}\|^2 \\
&\quad - 2\gamma(1-\theta) [f(x^{k,h}) - f(x^*)] \\
&\stackrel{(20)}{=} -\frac{\gamma\mu(1-\theta)}{2} \|\tilde{x}^{k,h} - x^*\|^2 \\
&\quad + \gamma^3\mu(1-\theta) \|g^{k,h}\|^2 \\
&\quad - 2\gamma(1-\theta) [f(x^{k,h}) - f(x^*)].
\end{aligned}$$

1104 Substituting this estimates into (24), we obtain

$$\begin{aligned}
\mathbb{E}_{\xi_m^{k,h}} \|\tilde{x}^{k,h+1} - x^*\|^2 &\leq \|\tilde{x}^{k,h} - x^*\|^2 + \theta \|\tilde{x}^{k,h} - x^*\|^2 + 2\gamma^2\theta \|g^{k-1, H^{k-1}}\|^2 \\
&\quad + \gamma^2(1-\theta)(1+\gamma\mu) \|g^{k,h}\|^2 + 3\gamma^2(1-\theta) \|\nabla f(x^{k,h})\|^2 \\
&\quad - \frac{\gamma\mu(1-\theta)}{2} \|\tilde{x}^{k,h} - x^*\|^2 - 2\gamma(1-\theta) [f(x^{k,h}) - f(x^*)] \\
&\quad + \frac{2\gamma^2(1-\theta)\sigma^2}{M} \\
&= \left(1 - \frac{\gamma\mu(1-\theta)}{2} + \theta\right) \|\tilde{x}^{k,h} - x^*\|^2 + 2\gamma^2\theta \|g^{k-1, H^{k-1}}\|^2 \\
&\quad + \gamma^2(1-\theta)(1+\gamma\mu) \|g^{k,h}\|^2 + 3\gamma^2(1-\theta) \|\nabla f(x^{k,h})\|^2 \\
&\quad - 2\gamma(1-\theta) [f(x^{k,h}) - f(x^*)] + \frac{2\gamma^2(1-\theta)\sigma^2}{M}. \tag{25}
\end{aligned}$$

1105 Let us choose $\theta \leq \frac{\gamma\mu}{4}$ and $\gamma \leq \frac{1}{L}$. Then, $\left(1 - \frac{\gamma\mu(1-\theta)}{2} + \theta\right) \leq \left(1 - \frac{3\gamma\mu}{8} + \frac{\gamma\mu}{4}\right) = \left(1 - \frac{\gamma\mu}{8}\right)$. In
1106 this way, (25) transforms to

$$\begin{aligned}
\mathbb{E}_{\xi_m^{k,h}} \|\tilde{x}^{k,h+1} - x^*\|^2 &\leq \left(1 - \frac{\gamma\mu}{8}\right) \|\tilde{x}^{k,h} - x^*\|^2 + 2\gamma^2\theta \|g^{k-1, H^{k-1}}\|^2 \\
&\quad + 2\gamma^2(1-\theta) \|g^{k,h}\|^2 + 3\gamma^2(1-\theta) \|\nabla f(x^{k,h})\|^2 \\
&\quad - 2\gamma(1-\theta) [f(x^{k,h}) - f(x^*)] + \frac{2\gamma^2(1-\theta)\sigma^2}{M}. \tag{26}
\end{aligned}$$

1107 Next we estimate

$$3\gamma^2(1-\theta) \|\nabla f(x^{k,h})\|^2 \stackrel{\text{(Lip)}}{\leq} 6\gamma^2L(1-\theta) [f(x^{k,h}) - f(x^*)]$$

1108 and combine with (25):

$$\begin{aligned}
\mathbb{E}_{\xi_m^{k,h}} \|\tilde{x}^{k,h+1} - x^*\|^2 &\leq \left(1 - \frac{\gamma\mu}{8}\right) \|\tilde{x}^{k,h} - x^*\|^2 + 2\gamma^2\theta \|g^{k-1,H^{k-1}}\|^2 \\
&\quad + 2\gamma^2(1-\theta) \|g^{k,h}\|^2 \\
&\quad - 2\gamma(1-\theta)(1-3\gamma L) [f(x^{k,h}) - f(x^*)] \\
&\quad + \frac{2\gamma^2(1-\theta)\sigma^2}{M}.
\end{aligned}$$

1109 By choosing $\gamma \leq \frac{1}{6L}$ we can simplify as

$$\begin{aligned}
\mathbb{E}_{\xi_m^{k,h}} \|\tilde{x}^{k,h+1} - x^*\|^2 &\leq \left(1 - \frac{\gamma\mu}{8}\right) \|\tilde{x}^{k,h} - x^*\|^2 + 2\gamma^2\theta \|g^{k-1,H^{k-1}}\|^2 \\
&\quad + 2\gamma^2(1-\theta) \|g^{k,h}\|^2 - \gamma(1-\theta) [f(x^{k,h}) - f(x^*)] \\
&\quad + \frac{2\gamma^2(1-\theta)\sigma^2}{M}.
\end{aligned}$$

1110 Now we put $h = H^k - 1$ and take additional expectations to obtain

$$\begin{aligned}
\mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|\tilde{x}^{k,H^k} - x^*\|^2 &\leq \left(1 - \frac{\gamma\mu}{8}\right) \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|\tilde{x}^{k,H^k-1} - x^*\|^2 \\
&\quad + 2\gamma^2\theta \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1-1}} \|g^{k-1,H^{k-1}}\|^2 \\
&\quad + 2\gamma^2(1-\theta) \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|g^{k,H^k-1}\|^2 \\
&\quad - \gamma(1-\theta) \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} [f(x^{k,H^k-1}) - f(x^*)] \\
&\quad + \frac{2\gamma^2(1-\theta)\sigma^2}{M}.
\end{aligned}$$

1111 We take expectation with respect to H^{k-1} and H^k , and apply Lemma 2:

$$\begin{aligned}
&\mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|\tilde{x}^{k,H^k} - x^*\|^2 \\
&\leq p \left(1 - \frac{\gamma\mu}{8}\right) \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1-1}} \|\tilde{x}^{k,0} - x^*\|^2 \\
&\quad + (1-p) \left(1 - \frac{\gamma\mu}{8}\right) \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|\tilde{x}^{k,H^k} - x^*\|^2 \\
&\quad + 2\gamma^2\theta \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1-1}} \|g^{k-1,H^{k-1}}\|^2 \\
&\quad + 2\gamma^2(1-\theta)(1-p) \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|g^{k,H^k}\|^2 \\
&\quad + 2\gamma^2(1-\theta)p \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1-1}} \underbrace{\|g^{k,0}\|^2}_{=0} \\
&\quad - \gamma(1-p)(1-\theta) \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} [f(x^{k,H^k}) - f(x^*)] \\
&\quad - \gamma p(1-\theta) \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1-1}} [f(x^{k,0}) - f(x^*)] \\
&\quad + \frac{2\gamma^2(1-\theta)\sigma^2}{M}.
\end{aligned}$$

1112 We rearrange terms and use that H^k and $\{\xi_m^{k-1,h}\}_{h=0}^{H^{k-1}-1}$ are independent stochastic values:

$$p \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|\tilde{x}^{k,H^k} - x^*\|^2$$

$$\begin{aligned}
&\leq p \left(1 - \frac{\gamma\mu}{8}\right) \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \|\tilde{x}^{k,0} - x^*\|^2 \\
&\quad + 2\gamma^2 \theta \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \|g^{k-1,H^{k-1}}\|^2 \\
&\quad + 2\gamma^2 (1-p)(1-\theta) \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|g^{k,H^k}\|^2 \\
&\quad - \gamma(1-p)(1-\theta) \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} [f(x^{k,H^k}) - f(x^*)] \\
&\quad - \gamma p(1-\theta) \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} [f(x^{k,0}) - f(x^*)] \\
&\quad + \frac{2\gamma^2(1-\theta)\sigma^2}{M}.
\end{aligned} \tag{27}$$

1113 We use Lemma 1 to estimate $\|g^{k,H^k}\|^2$ and $\|g^{k-1,H^{k-1}}\|^2$. We obtain

$$\begin{aligned}
\mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|g^{k,H^k}\|^2 &\leq \frac{24(1-\theta)^2\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^k} \|\nabla f(x^{k,H^k})\|^2 + \frac{48(1-\theta)^2\alpha\delta_2}{p^2} \\
&\quad + \frac{24(1-\theta)^2\alpha\sigma^2}{Mp^2}.
\end{aligned} \tag{28}$$

1114 As for $\|g^{k-1,H^{k-1}}\|^2$, we have

$$\begin{aligned}
&\mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \|g^{k-1,H^{k-1}}\|^2 \\
&\leq \frac{24(1-\theta)^2\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \|\nabla f(x^{k-1,H^{k-1}})\|^2 + \frac{48(1-\theta)^2\alpha\delta_2}{p^2} + \frac{24(1-\theta)^2\alpha\sigma^2}{Mp^2} \\
&\stackrel{\text{(CS)}}{\leq} \frac{48(1-\theta)\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \|\nabla f(x^{k-1,H^{k-1}}) - \nabla f(\tilde{x}^{k-1,H^{k-1}})\|^2 \\
&\quad + \frac{48(1-\theta)^2\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \|\nabla f(\tilde{x}^{k-1,H^{k-1}})\|^2 + \frac{48(1-\theta)^2\alpha\delta_2}{p^2} + \frac{24(1-\theta)^2\alpha\sigma^2}{Mp^2} \\
&\stackrel{\text{As. 1}}{\leq} \frac{48L^2(1-\theta)^2\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \|x^{k-1,H^{k-1}} - \tilde{x}^{k-1,H^{k-1}}\|^2 \\
&\quad + \frac{48(1-\theta)^2\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \|\nabla f(x^{k,0})\|^2 + \frac{48(1-\theta)^2\alpha\delta_2}{p^2} + \frac{24(1-\theta)^2\alpha\sigma^2}{Mp^2} \\
&\stackrel{\text{(20)}}{=} \frac{48\gamma^2 L^2(1-\theta)^2\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \|g^{k-1,H^{k-1}}\|^2 \\
&\quad + \frac{48(1-\theta)^2\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \|\nabla f(x^{k,0})\|^2 + \frac{48(1-\theta)^2\alpha\delta_2}{p^2} + \frac{24(1-\theta)^2\alpha\sigma^2}{Mp^2}.
\end{aligned}$$

1115 We choose $\gamma \leq \frac{p}{96L\sqrt{\alpha}\sqrt{\delta_1+1}}$. Moreover, we take additional expectations and again use that H^{k-1}

1116 and $\{\xi_m^{k-1,h}\}_{h=0}^{H^{k-1}-1}$ are independent stochastic values:

$$\begin{aligned}
&\mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \|g^{k-1,H^{k-1}}\|^2 \\
&\leq \frac{96(1-\theta)^2\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \|\nabla f(x^{k,0})\|^2 + \frac{96(1-\theta)^2\alpha\delta_2}{p^2} \\
&\quad + \frac{48(1-\theta)^2\alpha\sigma^2}{Mp^2}.
\end{aligned} \tag{29}$$

1117 Applying (Lip) to (28), (29) and substituting it to (27), we get

$$p \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|\tilde{x}^{k,H^k} - x^*\|^2$$

$$\begin{aligned}
&\leq p \left(1 - \frac{\gamma\mu}{8}\right) \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \|\tilde{x}^{k,0} - x^*\|^2 \\
&\quad - \gamma(1-p)(1-\theta) \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} [f(x^{k,H^k}) - f(x^*)] \\
&\quad - \gamma p(1-\theta) \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} [f(x^{k,0}) - f(x^*)] \\
&\quad + \frac{96\gamma^2 L(1-p)(1-\theta)^3 \alpha(\delta_1 + 1)}{p^2} \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H-1}} \mathbb{E}_{H^k} [f(x^{k,H^k}) - f(x^*)] \\
&\quad + \frac{96\gamma^2(1-p)(1-\theta)^3 \alpha \delta_2}{p^2} + \frac{48\gamma^2(1-p)(1-\theta)^3 \alpha \sigma^2}{Mp^2} \\
&\quad + \frac{384\gamma^2 L\theta(1-\theta)^2 \alpha(\delta_1 + 1)}{p^2} \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H-1}} [f(x^{k,0}) - f(x^*)] \\
&\quad + \frac{192\gamma^2 \theta(1-\theta)^2 \alpha \delta_2}{p^2} + \frac{96\gamma^2 \theta(1-\theta)^2 \alpha \sigma^2}{Mp^2} + \frac{2\gamma^2(1-\theta)\sigma^2}{M}.
\end{aligned}$$

1118 We take the full expectation, then use a law of expectation and rearrange terms:

$$\begin{aligned}
p\mathbb{E} \|\tilde{x}^{k,H^k} - x^*\|^2 &\leq p \left(1 - \frac{\gamma\mu}{8}\right) \mathbb{E} \|\tilde{x}^{k,0} - x^*\|^2 \\
&\quad - \gamma(1-p)(1-\theta) \left(1 - \frac{96\gamma L(1-\theta)^2 \alpha(\delta_1 + 1)}{p^2}\right) \cdot \\
&\quad \cdot \mathbb{E} [f(x^{k,H^k}) - f(x^*)] \\
&\quad - \gamma p(1-\theta) \left(1 - \frac{384\gamma L\theta(1-\theta)\alpha(\delta_1 + 1)}{p^3}\right) \mathbb{E} [f(x^{k,0}) - f(x^*)] \\
&\quad + \frac{96\gamma^2(1-p)(1-\theta)^3 \alpha \delta_2}{p^2} + \frac{192\gamma^2 \theta(1-\theta)^2 \alpha \delta_2}{p^2} \\
&\quad + \frac{48\gamma^2(1-p)(1-\theta)^3 \alpha \sigma^2}{Mp^2} + \frac{96\gamma^2 \theta(1-\theta)^2 \alpha \sigma^2}{Mp^2} \\
&\quad + \frac{2\gamma^2(1-\theta)\sigma^2}{M}.
\end{aligned}$$

1119 We choose $\theta \leq \frac{p\gamma\mu}{4}$ and $\gamma \leq \frac{p^2}{96L\alpha(\delta_1+1)}$. Note that all previous transitions hold even with larger
1120 choice of θ and γ , consequently this choice is correct. In that way, we obtain

$$\begin{aligned}
\mathbb{E} \|\tilde{x}^{k,H^k} - x^*\|^2 &\leq \left(1 - \frac{\gamma\mu}{8}\right) \mathbb{E} \|\tilde{x}^{k,0} - x^*\|^2 + \frac{96\gamma^2 \alpha \delta_2}{p^3} + \frac{48\gamma^3 \mu \alpha \delta_2}{p^2} \\
&\quad + \frac{48\gamma^2 \alpha \sigma^2}{Mp^3} + \frac{24\gamma^3 \mu \alpha \sigma^2}{Mp^2} + \frac{2\gamma^2 \sigma^2}{Mp}.
\end{aligned}$$

1121 Note that $\tilde{x}^{k,H^k} = x^{k,H^k} - \gamma g^{k,H^k} = x^{k+1,0}$ and $\tilde{x}^{k,0} = x^{k,0}$. Thus,

$$\mathbb{E} \|x^{k+1,0} - x^*\|^2 \leq \left(1 - \frac{\gamma\mu}{8}\right) \mathbb{E} \|x^{k,0} - x^*\|^2 + \frac{\gamma^2 \alpha}{p^3} \left(144\delta_2 + \frac{74\sigma^2}{M}\right).$$

1122 It remains for us to take into account going into recursion over all epochs and claim the result of the
1123 theorem:

$$\begin{aligned}
\mathbb{E} \|x^{K,0} - x^*\|^2 &\leq \left(1 - \frac{\gamma\mu}{8}\right)^K \|x^{0,0} - x^*\|^2 + \frac{\gamma^2 \alpha}{p^3} \left(144\delta_2 + \frac{74\sigma^2}{M}\right) \sum_{k=0}^K \left(1 - \frac{\gamma\mu}{8}\right)^k \\
&\leq \left(1 - \frac{\gamma\mu}{8}\right)^K \|x^{0,0} - x^*\|^2 + \frac{8\gamma \alpha}{\mu p^3} \left(144\delta_2 + \frac{74\sigma^2}{M}\right).
\end{aligned}$$

1125 **Corollary 7 (Corollary 2).** Under conditions of Theorem 2 Algorithm 1 with fixed rules $\widehat{\mathcal{R}} \equiv \widetilde{\mathcal{R}} \equiv \mathcal{R}$
 1126 needs

$$\begin{aligned} & \widetilde{\mathcal{O}} \left(\left(\frac{M}{C} \right)^2 \left(\frac{L}{\mu} \alpha \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{C} \frac{\alpha \delta_2}{\mu^2 \varepsilon} + \frac{\alpha \sigma^2}{\mu^2 C \varepsilon} \right) \right) \text{ epochs and} \\ & \widetilde{\mathcal{O}} \left(M \left(\frac{M}{C} \right)^2 \left(\frac{L}{\mu} \alpha \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{C} \frac{\alpha \delta_2}{\mu^2 \varepsilon} + \frac{\alpha \sigma^2}{\mu^2 C \varepsilon} \right) \right) \text{ number of devices communications} \end{aligned}$$

1127 to reach ε -accuracy, where $\varepsilon^2 = \mathbb{E} \|x^{K,0} - x^*\|^2$ and C is the number devices participating in each
 1128 epoch.

1129 *Proof.* Using the result of Theorem 2, we choose

$$\gamma \leq \min \left\{ \frac{p^2}{96L\alpha(\delta_1 + 1)}, \frac{8 \log \left(\max \left\{ 2, \frac{\mu^2 M p^3 \|x^{0,0} - x^*\|^2 K}{4736\alpha\sigma^2}, \frac{\mu^2 p^3 \|x^{0,0} - x^*\|^2 K}{9216\alpha\delta_2} \right\} \right)}{\mu K} \right\}$$

1130 Thus, we need $\widetilde{\mathcal{O}} \left(\frac{L\alpha\delta_1}{\mu p^2} \log \left(\frac{1}{\varepsilon} \right) + \frac{\alpha\delta_2}{\mu^2 p^3 \varepsilon} + \frac{\alpha\sigma^2}{\mu^2 M p^3 \varepsilon} \right)$ epochs to reach ε -accuracy, where $\varepsilon^2 =$
 1131 $\mathbb{E} \|x^{K,0} - x^*\|^2$. Since the length of the epoch $H \in \text{Geom}(p)$, Algorithm 1 requires
 1132 $\widetilde{\mathcal{O}} \left(\frac{L\alpha\delta_1}{\mu p^3} \log \left(\frac{1}{\varepsilon} \right) + \frac{\alpha\delta_2}{\mu^2 p^4 \varepsilon} + \frac{\alpha\sigma^2}{\mu^2 M p^4 \varepsilon} \right)$ communication rounds. Next we mention that at each
 1133 communication round we communicate with C devices, thus, number of communications is
 1134 $\widetilde{\mathcal{O}} \left(C \left(\frac{L\alpha\delta_1}{\mu p^3} \log \left(\frac{1}{\varepsilon} \right) + \frac{\alpha\delta_2}{\mu^2 p^4 \varepsilon} + \frac{\alpha\sigma^2}{\mu^2 M p^4 \varepsilon} \right) \right)$. Taking, $p = \frac{C}{M}$, we have the result of the corollary.
 1135 The choice of p is motivated by the fact that we perform $\frac{1}{p}C + M$ communications per-epoch, and
 1136 established p is the minimal, which delivers $\mathcal{O}(M)$ communications at each epoch. This is also the
 1137 reason for the additional factor M in the estimate on communications. □

1138 **Corollary 8.** Under conditions of Theorem 2 Algorithm 1 needs

$$\begin{aligned} & \widetilde{\mathcal{O}} \left(\left(\frac{M}{\min_{k,h} C^{k,h}} \right)^2 \left(\frac{L}{\mu} \alpha \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{\min_{k,h} C^{k,h}} \frac{\alpha \delta_2}{\mu^2 \varepsilon} + \frac{\alpha \sigma^2}{\mu^2 \min_{k,h} C^{k,h} \varepsilon} \right) \right) \text{ epochs and} \\ & \widetilde{\mathcal{O}} \left(M \left(\frac{M}{\min_{k,h} C^{k,h}} \right)^3 \left(\frac{L}{\mu} \alpha \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{\min_{k,h} C^{k,h}} \frac{\alpha \delta_2}{\mu^2 \varepsilon} + \frac{\alpha \sigma^2}{\mu^2 \min_{k,h} C^{k,h} \varepsilon} \right) \right) \end{aligned}$$

1139 number of devices communications to reach ε -accuracy, where $\varepsilon^2 = \mathbb{E} \|x^{K,0} - x^*\|^2$ and $C^{k,h}$ is
 1140 the number of devices participating in k -th iteration in h -th epoch.

1141 *Proof.* Using the result of Theorem 2, we choose

$$\gamma \leq \min \left\{ \frac{p^2}{96L\alpha(\delta_1 + 1)}, \frac{8 \log \left(\max \left\{ 2, \frac{\mu^2 M p^3 \|x^{0,0} - x^*\|^2 K}{4736\alpha\sigma^2}, \frac{\mu^2 p^3 \|x^{0,0} - x^*\|^2 K}{9216\alpha\delta_2} \right\} \right)}{\mu K} \right\}$$

1142 Thus, we need $\widetilde{\mathcal{O}} \left(\frac{L\alpha\delta_1}{\mu p^2} \log \left(\frac{1}{\varepsilon} \right) + \frac{\alpha\delta_2}{\mu^2 p^3 \varepsilon} + \frac{\alpha\sigma^2}{\mu^2 M p^3 \varepsilon} \right)$ epochs to reach ε -accuracy, where $\varepsilon^2 =$
 1143 $\mathbb{E} \|x^{K,0} - x^*\|^2$. Since the length of the epoch $H \in \text{Geom}(p)$, Algorithm 1 requires
 1144 $\widetilde{\mathcal{O}} \left(\frac{L\alpha\delta_1}{\mu p^3} \log \left(\frac{1}{\varepsilon} \right) + \frac{\alpha\delta_2}{\mu^2 p^4 \varepsilon} + \frac{\alpha\sigma^2}{\mu^2 M p^4 \varepsilon} \right)$ communication rounds. Next we mention that at each
 1145 communication round we communicate with $C^{k,h}$ devices, thus, number of communications is
 1146 $\widetilde{\mathcal{O}} \left(\max_{k,h} C^{k,h} \left(\frac{L\alpha\delta_1}{\mu p^3} \log \left(\frac{1}{\varepsilon} \right) + \frac{\alpha\delta_2}{\mu^2 p^4 \varepsilon} + \frac{\alpha\sigma^2}{\mu^2 M p^4 \varepsilon} \right) \right)$. Taking, $p = \frac{\min_{k,h} C^{k,h}}{M}$, we have the result of the
 1147 corollary. The choice of p is motivated by the fact that we perform $\frac{1}{p} \max_{k,h} C^{k,h} + M$ communications

per-epoch, and established p is the minimal, which delivers $\mathcal{O}\left(M \frac{M}{\min_{k,h} C^{k,h}}\right)$ communications at each epoch while guarantee the epoch is executed (if we take $p = \frac{\max_{k,h} C^{k,h}}{M}$, we can meet $p = 1$). This is also the reason for the additional factor $M \frac{M}{\min_{k,h} C^{k,h}}$ in the estimate on communications. \square

Remark 2. Considering fixed rules $\widehat{\mathcal{R}} \equiv \widetilde{\mathcal{R}} \equiv \mathcal{R}$, we have $\widetilde{\mathcal{O}}\left(M \left(\frac{M}{C}\right)^2 \left(\frac{L}{\mu} \delta_1 \log\left(\frac{1}{\varepsilon}\right) + \frac{M}{C} \frac{\delta_2}{\mu^2 \varepsilon} + \frac{\sigma^2}{\mu^2 C \varepsilon}\right)\right)$ and $\widetilde{\mathcal{O}}\left(M^2 \left(\frac{M}{C}\right)^2 \left(\frac{L}{\mu} \delta_1 \log\left(\frac{1}{\varepsilon}\right) + \frac{M}{C} \frac{\delta_2}{\mu^2 \varepsilon} + \frac{\sigma^2}{\mu^2 C \varepsilon}\right)\right)$ number of devices communications with regularizing parameter $\alpha = 1$ and $\alpha = M$ respectively. Considering various rules, best case with regularizing coefficient $\alpha = 1$ gives us $\widetilde{\mathcal{O}}\left(M \left(\frac{M}{\min_{k,h} C^{k,h}}\right)^3 \left(\frac{L}{\mu} \delta_1 \log\left(\frac{1}{\varepsilon}\right) + \frac{M}{\min_{k,h} C^{k,h}} \frac{\delta_2}{\mu^2 \varepsilon} + \frac{\sigma^2}{\mu^2 \min_{k,h} C^{k,h} \varepsilon}\right)\right)$ and worst case $\alpha = M$ gives us $\widetilde{\mathcal{O}}\left(M^2 \left(\frac{M}{\min_{k,h} C^{k,h}}\right)^3 \left(\frac{L}{\mu} \delta_1 \log\left(\frac{1}{\varepsilon}\right) + \frac{M}{\min_{k,h} C^{k,h}} \frac{\delta_2}{\mu^2 \varepsilon} + \frac{\sigma^2}{\mu^2 \min_{k,h} C^{k,h} \varepsilon}\right)\right)$ number of devices communications.

E Proofs for Algorithm 2

Lemma 4. Suppose Assumptions 3, 4 hold. Then for Algorithm 2 it implies that

$$\begin{aligned} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k,0}} \mathbb{E}_{\eta_m^{k,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \mathbb{E}_{\eta_m^{k,H^k-1}} \|g^{k,H^k}\|^2 &\leq \frac{96(1-\theta)^2 \alpha (\delta_1 + 1)}{p^2 \min_{1 \leq m \leq M} q_m} \mathbb{E}_{H^k} \|\nabla f(x^{k,H^k})\|^2 \\ &\quad + \frac{192(1-\theta)^2 \alpha \delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{96(1-\theta)^2 \alpha \sigma^2}{p^2 \min_{1 \leq m \leq M} q_m}. \end{aligned}$$

Proof. Let us start with the following estimate:

$$\begin{aligned} \|g^{k,h+1}\|^2 &= \left\| g^{k,h} + (1-\theta) \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \left(\frac{1}{M} - \hat{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\ &\stackrel{(\text{Fen})}{\leq} (1+c) \|g^{k,h}\|^2 \\ &\quad + \left(1 + \frac{1}{c}\right) (1-\theta)^2 \left\| \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \left(\frac{1}{M} - \hat{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2, \quad (30) \end{aligned}$$

where c is defined below. Let us estimate the last term and obtain

$$\begin{aligned} &\left\| \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \left(\frac{1}{M} - \hat{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\ &\stackrel{(\text{CS})}{\leq} 2 \left\| \sum_{m=1}^M \left(\frac{\eta_m^{k,h}}{q_m} - 1 \right) \left(\frac{1}{M} - \hat{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\ &\quad + 2 \left\| \sum_{m=1}^M \left(\frac{1}{M} - \hat{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\ &\stackrel{(\text{CS})}{\leq} 2 \sum_{m=1}^M \left(\frac{\eta_m^{k,h}}{q_m} - 1 \right)^2 \left(\frac{1}{M} - \hat{\pi}_m^{k,h} \right)^2 \sum_{m=1}^M \|\nabla f_m(x^{k,h}, \xi_m^{k,h})\|^2 \\ &\quad + 2 \left\| \sum_{m=1}^M \left(\frac{1}{M} - \hat{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2. \end{aligned}$$

1163 We pay attention to the first term. Using $\eta_m^{k,h} \sim \mathcal{B}(q_m)$,

$$\mathbb{E}_{\eta_m^{k,h}} \left(\frac{\eta_m^{k,h}}{q_m} - 1 \right)^2 = \frac{\mathbb{E}_{\eta_m^{k,h}} (\eta_m^{k,h} - q_m)^2}{(q_m)^2} \leq \frac{\sigma_\eta^2}{(q_m)^2} = \frac{1 - q_m}{q_m} \leq \frac{1}{q_m}.$$

1164 In that way,

$$\begin{aligned} & \mathbb{E}_{\eta_m^{k,h}} \left\| \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \left(\frac{1}{M} - \hat{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\ & \leq \frac{2}{\min_{1 \leq m \leq M} q_m} \sum_{m=1}^M \left(\frac{1}{M} - \hat{\pi}_m^{k,h} \right)^2 \sum_{m=1}^M \|\nabla f_m(x^{k,h}, \xi_m^{k,h})\|^2 \\ & \quad + 2 \left\| \sum_{m=1}^M \left(\frac{1}{M} - \hat{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2. \end{aligned} \quad (31)$$

1165 We obtained an estimate for the second term in Lemma 3 in (7):

$$\mathbb{E}_{\xi_m^{k,h}} \left\| \sum_{m=1}^M \left(\frac{1}{M} - \hat{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}) \right\|^2 \leq 4\alpha(\delta_1 + 1) \|\nabla f(x^{k,h})\|^2 + 4\alpha\delta_2 + \frac{2\alpha\sigma^2}{M}.$$

1166 Moreover, in (6) we found out

$$\sum_{m=1}^M \left(\frac{1}{M} - \hat{\pi}_m^{k,h} \right)^2 \leq \frac{\alpha}{M}.$$

1167 Combining this estimates with (31),

$$\begin{aligned} & \mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} \left\| \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \left(\frac{1}{M} - \hat{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\ & \leq \frac{2\alpha}{\min_{1 \leq m \leq M} q_m M} \sum_{m=1}^M \mathbb{E}_{\xi_m^{k,h}} \|\nabla f_m(x^{k,h}, \xi_m^{k,h})\|^2 \\ & \quad + 8\alpha(\delta_1 + 1) \|\nabla f(x^{k,h})\|^2 + 8\alpha\delta_2 + \frac{4\alpha\sigma^2}{M} \\ & \stackrel{(CS)}{\leq} \frac{4\alpha}{\min_{1 \leq m \leq M} q_m M} \sum_{m=1}^M \mathbb{E}_{\xi_m^{k,h}} \|\nabla f_m(x^{k,h})\|^2 \\ & \quad + \frac{4\alpha}{\min_{1 \leq m \leq M} q_m M} \sum_{m=1}^M \mathbb{E}_{\xi_m^{k,h}} \|\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})\|^2 \\ & \quad + 8\alpha(\delta_1 + 1) \|\nabla f(x^{k,h})\|^2 + 8\alpha\delta_2 + \frac{4\alpha\sigma^2}{M} \\ & \stackrel{As. 4}{\leq} \frac{4\alpha}{\min_{1 \leq m \leq M} q_m M} \sum_{m=1}^M \|\nabla f_m(x^{k,h})\|^2 + \frac{4\alpha\sigma^2}{\min_{1 \leq m \leq M} q_m} \\ & \quad + 8\alpha(\delta_1 + 1) \|\nabla f(x^{k,h})\|^2 + 8\alpha\delta_2 + \frac{4\alpha\sigma^2}{M} \\ & \stackrel{(CS)}{\leq} \frac{8\alpha}{\min_{1 \leq m \leq M} q_m M} \sum_{m=1}^M \|\nabla f(x^{k,h})\|^2 + \frac{8\alpha}{\min_{1 \leq m \leq M} q_m M} \sum_{m=1}^M \|\nabla f_m(x^{k,h}) - \nabla f(x^{k,h})\|^2 \end{aligned}$$

$$\begin{aligned}
& + 8\alpha(\delta_1 + 1) \|\nabla f(x^{k,h})\|^2 + 8\alpha\delta_2 + \frac{8\alpha\sigma^2}{\min_{1 \leq m \leq M} q_m} \\
& \stackrel{\text{As. 3}}{\leq} \frac{8\alpha(\delta_1 + 1)}{\min_{1 \leq m \leq M} q_m} \|\nabla f(x^{k,h})\|^2 + 8\alpha(\delta_1 + 1) \|\nabla f(x^{k,h})\|^2 + \frac{8\alpha\delta_2}{\min_{1 \leq m \leq M} q_m} \\
& + 8\alpha\delta_2 + \frac{8\alpha\sigma^2}{\min_{1 \leq m \leq M} q_m} \\
& \leq \frac{16\alpha(\delta_1 + 1)}{\min_{1 \leq m \leq M} q_m} \|\nabla f(x^{k,h})\|^2 + \frac{16\alpha\delta_2}{\min_{1 \leq m \leq M} q_m} + \frac{8\alpha\sigma^2}{\min_{1 \leq m \leq M} q_m}.
\end{aligned}$$

1168 Substituting this estimate into (30), we have

$$\begin{aligned}
\mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} \|g^{k,h+1}\|^2 & \leq (1+c) \|g^{k,h}\|^2 + 16 \left(1 + \frac{1}{c}\right) (1-\theta)^2 \frac{\alpha(\delta_1 + 1)}{\min_{1 \leq m \leq M} q_m} \|\nabla f(x^{k,h})\|^2 \\
& + 16 \left(1 + \frac{1}{c}\right) (1-\theta)^2 \frac{\alpha\delta_2}{\min_{1 \leq m \leq M} q_m} \\
& + 8 \left(1 + \frac{1}{c}\right) (1-\theta)^2 \frac{\alpha\sigma^2}{\min_{1 \leq m \leq M} q_m}.
\end{aligned}$$

1169 Enrolling a recursion, we get

$$\begin{aligned}
& \mathbb{E}_{\xi_m^{k,0}} \mathbb{E}_{\eta_m^{k,0}} \dots \mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} \|g^{k,h+1}\|^2 \\
& \leq 16 \left(1 + \frac{1}{c}\right) (1-\theta)^2 \frac{\alpha(\delta_1 + 1)}{\min_{1 \leq m \leq M} q_m} \sum_{i=0}^h (1+c)^{h-i} \|\nabla f(x^{k,i})\|^2 \\
& + 16 \left(1 + \frac{1}{c}\right) (1-\theta)^2 \frac{\alpha\delta_2}{\min_{1 \leq m \leq M} q_m} \sum_{i=0}^h (1+c)^{h-i} \\
& + 8 \left(1 + \frac{1}{c}\right) (1-\theta)^2 \frac{\alpha\sigma^2}{\min_{1 \leq m \leq M} q_m} \sum_{i=0}^h (1+c)^{h-i}. \tag{32}
\end{aligned}$$

1170 Next, choosing $c = \frac{p}{2}$, taking exception on H^k and applying (9), (10), (11) from Lemma 3, we obtain
1171 the result of the lemma:

$$\begin{aligned}
\mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k,0}} \mathbb{E}_{\eta_m^{k,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \mathbb{E}_{\eta_m^{k,H^k-1}} \|g^{k,H^k}\|^2 & \leq \frac{96(1-\theta)^2 \alpha(\delta_1 + 1)}{p^2 \min_{1 \leq m \leq M} q_m} \mathbb{E}_{H^k} \|\nabla f(x^{k,H^k})\|^2 \\
& + \frac{192(1-\theta)^2 \alpha\delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{96(1-\theta)^2 \alpha\sigma^2}{p^2 \min_{1 \leq m \leq M} q_m}.
\end{aligned}$$

1172

□

1173 E.1 Proof for non-convex setting

1174 **Theorem 5.** Suppose Assumptions 1, 2(a), 3, 4 hold. Then for Algorithm 2 with $\theta \leq \frac{\gamma L p^2}{2}$ and

1175 $\gamma \leq \frac{p \min_{1 \leq m \leq M} q_m}{768 L \alpha(\delta_1 + 1)}$ it implies that

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(x^{k,0})\|^2 \leq \frac{16(f(x^{0,0}) - f(x^*))}{\gamma K}$$

$$\begin{aligned}
& + \frac{1536\gamma^2 L^2 \alpha \delta_2}{p^3 \min_{1 \leq m \leq M} q_m} + \frac{3200\gamma L \alpha \delta_2}{p \min_{1 \leq m \leq M} q_m} \\
& + \frac{768\gamma^2 L^2 \alpha \sigma^2}{p^3 \min_{1 \leq m \leq M} q_m} + \frac{1600\gamma L \alpha \sigma^2}{p \min_{1 \leq m \leq M} q_m}.
\end{aligned}$$

1176 *Proof.* We start with the definition of virtual sequence:

$$\tilde{x}^{k,h} = x^{k,h} - \gamma \sum_{m=1}^M g_m^{k,h} = x^{k,h} - \gamma g^{k,h}. \quad (33)$$

1177 It is followed by

$$\begin{aligned}
\tilde{x}^{k,h+1} &= x^{k,h+1} - \gamma \sum_{m=1}^M g_m^{k,h+1} \\
&= x^{k,h} - \gamma \left[(1-\theta) \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \hat{\pi}_m^{k,h} \nabla f_m(x^{k,h}, \xi_m^{k,h}) + \theta g^{k-1, H^{k-1}} \right] \\
&\quad - \gamma \sum_{m=1}^M g_m^{k,h} - \gamma (1-\theta) \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \left(\frac{1}{M} - \hat{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \\
&= \tilde{x}^{k,h} - \gamma \left[(1-\theta) \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) + \theta g^{k-1, H^{k-1}} \right]. \quad (34)
\end{aligned}$$

1178 Assumption 1 implies

$$\begin{aligned}
f(\tilde{x}^{k,h+1}) &\leq f(\tilde{x}^{k,h}) + \langle \nabla f(\tilde{x}^{k,h}), \tilde{x}^{k,h+1} - \tilde{x}^{k,h} \rangle + \frac{L}{2} \|\tilde{x}^{k,h+1} - \tilde{x}^{k,h}\|^2 \\
&\stackrel{(34)}{\leq} f(\tilde{x}^{k,h}) - \gamma \theta \langle \nabla f(\tilde{x}^{k,h}), g^{k-1, H^{k-1}} \rangle \\
&\quad - \gamma (1-\theta) \left\langle \nabla f(\tilde{x}^{k,h}), \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\
&\quad + \frac{\gamma^2 L (1-\theta)}{2} \left\| \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 + \frac{\gamma^2 L \theta}{2} \|g^{k-1, H^{k-1}}\|^2.
\end{aligned}$$

1179 Now we use that $\eta_m^{k,h} \sim \mathcal{B}(q_m)$. Consequently, $\mathbb{E} \eta_m^{k,h} = q_m$. Since $\eta_m^{k,h}$ is independent of
1180 $x^{k,h}, \tilde{x}^{k,h}, \xi_m^{k,h}, g^{k-1, H^{k-1}}$, we take the expectation and obtain

$$\begin{aligned}
\mathbb{E}_{\eta_m^{k,h}} [f(\tilde{x}^{k,h+1})] &\leq f(\tilde{x}^{k,h}) - \gamma \theta \langle \nabla f(\tilde{x}^{k,h}), g^{k-1, H^{k-1}} \rangle \\
&\quad - \gamma (1-\theta) \left\langle \nabla f(\tilde{x}^{k,h}), \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\
&\quad + \frac{\gamma^2 L (1-\theta)}{2} \mathbb{E}_{\eta_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\
&\quad + \frac{\gamma^2 L \theta}{2} \|g^{k-1, H^{k-1}}\|^2.
\end{aligned}$$

1181 We take the expectation over $\xi_m^{k,h}$. Mention that

$$\begin{aligned}
\tilde{x}^{k,h} &\stackrel{(33)}{=} x^{k,h} - \gamma g^{k,h} \\
&\stackrel{\text{Line 12}}{=} x^{k,h} - \gamma \left(g^{k,h-1} + (1-\theta) \sum_{m=1}^M \frac{\eta_m^{k,h-1}}{q_m} \left(\frac{1}{M} - \hat{\pi}_m^{k,h-1} \right) \nabla f(x^{k,h-1}, \xi_m^{k,h-1}) \right).
\end{aligned}$$

1182 Thus, $\tilde{x}^{k,h}$ and $\xi_m^{k,h}$ are independent. Analogously, $g^{k-1,H^{k-1}}$ and $\xi_m^{k,h}$ are independent. In this way,

$$\begin{aligned}
\mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} [f(\tilde{x}^{k,h+1})] &\leq f(\tilde{x}^{k,h}) - \gamma\theta \langle \nabla f(\tilde{x}^{k,h}), g^{k-1,H^{k-1}} \rangle \\
&\quad - \gamma(1-\theta) \langle \nabla f(\tilde{x}^{k,h}), \nabla f(x^{k,h}) \rangle \\
&\quad + \frac{\gamma^2 L(1-\theta)}{2} \mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\
&\quad + \frac{\gamma^2 L\theta}{2} \|g^{k-1,H^{k-1}}\|^2.
\end{aligned} \tag{35}$$

1183 Let us consider separately the following term:

$$\begin{aligned}
&\mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\
&\stackrel{\text{(CS)}}{\leq} 2\mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) - \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\
&\quad + 2\mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\
&= 2\mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \left(\frac{\eta_m^{k,h}}{q_m} - 1 \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\
&\quad + 2\mathbb{E}_{\xi_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\
&\stackrel{\text{(CS)}}{\leq} \frac{2}{M^2} \mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} \sum_{m=1}^M \left(\frac{\eta_m^{k,h}}{q_m} - 1 \right)^2 \sum_{m=1}^M \|\nabla f_m(x^{k,h}, \xi_m^{k,h})\|^2 \\
&\quad + 2\mathbb{E}_{\xi_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2.
\end{aligned} \tag{36}$$

1184 We pay attention to the first term. Using $\eta_m^{k,h} \sim \mathcal{B}(q_m)$,

$$\begin{aligned}
\sum_{m=1}^M \mathbb{E}_{\eta_m^{k,h}} \left(\frac{\eta_m^{k,h}}{q_m} - 1 \right)^2 &= \sum_{m=1}^M \frac{\mathbb{E}_{\eta_m^{k,h}} (\eta_m^{k,h} - q_m)^2}{(q_m)^2} \leq \sum_{m=1}^M \frac{\sigma_\eta^2}{(q_m)^2} \\
&= \sum_{m=1}^M \frac{1 - q_m}{q_m} \leq \frac{M}{\min_{1 \leq m \leq M} q_m}.
\end{aligned}$$

1185 Combining with (36),

$$\begin{aligned}
&\mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\
&\leq \frac{2}{M \min_{1 \leq m \leq M} q_m} \sum_{m=1}^M \mathbb{E}_{\xi_m^{k,h}} \|\nabla f_m(x^{k,h}, \xi_m^{k,h})\|^2 + 2\mathbb{E}_{\xi_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\
&\stackrel{\text{(CS)}}{\leq} \frac{4}{M \min_{1 \leq m \leq M} q_m} \sum_{m=1}^M \mathbb{E}_{\xi_m^{k,h}} \|\nabla f_m(x^{k,h}, \xi_m^{k,h})\|^2
\end{aligned}$$

$$\begin{aligned}
& \stackrel{\text{(CS)}}{\leq} \frac{8}{M \min_{1 \leq m \leq M} q_m} \sum_{m=1}^M \|\nabla f_m(x^{k,h})\|^2 \\
& + \frac{8}{M \min_{1 \leq m \leq M} q_m} \sum_{m=1}^M \mathbb{E}_{\xi_m^{k,h}} \|\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})\|^2 \\
& \stackrel{\text{As. 4}}{\leq} \frac{8}{M \min_{1 \leq m \leq M} q_m} \sum_{m=1}^M \|\nabla f_m(x^{k,h})\|^2 + \frac{8\sigma^2}{\min_{1 \leq m \leq M} q_m} \\
& \stackrel{\text{(CS)}}{\leq} \frac{16}{M \min_{1 \leq m \leq M} q_m} \sum_{m=1}^M \|\nabla f(x^{k,h})\|^2 \\
& + \frac{16}{M \min_{1 \leq m \leq M} q_m} \sum_{m=1}^M \|\nabla f_m(x^{k,h}) - \nabla f(x^{k,h})\|^2 + \frac{8\sigma^2}{\min_{1 \leq m \leq M} q_m} \\
& \stackrel{\text{As. 3}}{\leq} \frac{16(\delta_1 + 1)}{\min_{1 \leq m \leq M} q_m} \|\nabla f(x^{k,h})\|^2 + \frac{16\delta_2}{\min_{1 \leq m \leq M} q_m} + \frac{8\sigma^2}{\min_{1 \leq m \leq M} q_m}. \tag{37}
\end{aligned}$$

1186 We substitute this estimate into (35) to obtain

$$\begin{aligned}
\mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} [f(\tilde{x}^{k,h+1})] & \leq f(\tilde{x}^{k,h}) - \gamma\theta \langle \nabla f(\tilde{x}^{k,h}), g^{k-1, H^{k-1}} \rangle \\
& - \gamma(1-\theta) \langle \nabla f(\tilde{x}^{k,h}), \nabla f(x^{k,h}) \rangle \\
& + \frac{8\gamma^2 L(1-\theta)(\delta_1 + 1)}{\min_{1 \leq m \leq M} q_m} \|\nabla f(x^{k,h})\|^2 \\
& + \frac{\gamma^2 L\theta}{2} \|g^{k-1, H^{k-1}}\|^2 + \frac{8\gamma^2 L(1-\theta)\delta_2}{\min_{1 \leq m \leq M} q_m} \\
& + \frac{4\gamma^2 L(1-\theta)\sigma^2}{\min_{1 \leq m \leq M} q_m}. \tag{38}
\end{aligned}$$

1187 Let us estimate the scalar products separately.

$$\begin{aligned}
-\gamma(1-\theta) \langle \nabla f(\tilde{x}^{k,h}), \nabla f(x^{k,h}) \rangle & = -\frac{\gamma(1-\theta)}{2} \|\nabla f(\tilde{x}^{k,h})\|^2 - \frac{\gamma(1-\theta)}{2} \|\nabla f(x^{k,h})\|^2 \\
& + \frac{\gamma(1-\theta)}{2} \|\nabla f(\tilde{x}^{k,h}) - \nabla f(x^{k,h})\|^2 \\
& \stackrel{\text{As. 1}}{\leq} -\frac{\gamma(1-\theta)}{2} \|\nabla f(\tilde{x}^{k,h})\|^2 - \frac{\gamma(1-\theta)}{2} \|\nabla f(x^{k,h})\|^2 \\
& + \frac{\gamma L^2(1-\theta)}{2} \|\tilde{x}^{k,h} - x^{k,h}\|^2 \\
& \stackrel{(12)}{=} -\frac{\gamma(1-\theta)}{2} \|\nabla f(\tilde{x}^{k,h})\|^2 - \frac{\gamma(1-\theta)}{2} \|\nabla f(x^{k,h})\|^2 \\
& + \frac{\gamma^3 L^2(1-\theta)}{2} \|g^{k,h}\|^2, \\
-\gamma\theta \langle \nabla f(\tilde{x}^{k,h}), g^{k-1, H^{k-1}} \rangle & \stackrel{\text{(Fen)}}{\leq} \frac{\gamma\theta}{2} \|\nabla f(\tilde{x}^{k,h})\|^2 + \frac{\gamma\theta}{2} \|g^{k-1, H^{k-1}}\|^2.
\end{aligned}$$

1188 Combining it with (38),

$$\mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} [f(\tilde{x}^{k,h+1})] \leq f(\tilde{x}^{k,h}) - \frac{\gamma(1-\theta)}{2} \left(1 - \frac{16\gamma L(\delta_1 + 1)}{\min_{1 \leq m \leq M} q_m} \right) \|\nabla f(x^{k,h})\|^2$$

$$\begin{aligned}
& -\frac{\gamma(1-2\theta)}{2} \|\nabla f(\tilde{x}^{k,h})\|^2 + \frac{\gamma^3 L^2(1-\theta)}{2} \|g^{k,h}\|^2 \\
& + \frac{\gamma\theta(\gamma L+1)}{2} \|g^{k-1,H^{k-1}}\|^2 + \frac{8\gamma^2 L(1-\theta)\delta_2}{\min_{1 \leq m \leq M} q_m} + \frac{4\gamma^2 L(1-\theta)\sigma^2}{\min_{1 \leq m \leq M} q_m}.
\end{aligned}$$

1189 Choosing $\gamma \leq \frac{\min_{1 \leq m \leq M} q_m}{32L(\delta_1+1)}$ and $\theta \leq \frac{1}{2}$,

$$\begin{aligned}
\mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} [f(\tilde{x}^{k,h+1})] & \leq f(\tilde{x}^{k,h}) - \frac{\gamma(1-\theta)}{4} \|\nabla f(x^{k,h})\|^2 + \frac{\gamma^3 L^2(1-\theta)}{2} \|g^{k,h}\|^2 \\
& + \gamma\theta \|g^{k-1,H^{k-1}}\|^2 + \frac{8\gamma^2 L(1-\theta)\delta_2}{\min_{1 \leq m \leq M} q_m} + \frac{4\gamma^2 L(1-\theta)\sigma^2}{\min_{1 \leq m \leq M} q_m}.
\end{aligned}$$

1190 Now we put $h = H^k - 1$ and take additional expectations.

$$\begin{aligned}
& \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \mathbb{E}_{\eta_m^{k,H^k-1}} [f(\tilde{x}^{k,H^k})] \\
& \leq \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \mathbb{E}_{\eta_m^{k,H^k-1}} [f(\tilde{x}^{k,H^k-1})] \\
& \quad - \frac{\gamma(1-\theta)}{4} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \mathbb{E}_{\eta_m^{k,H^k-1}} \|\nabla f(x^{k,H^k-1})\|^2 \\
& \quad + \frac{\gamma^3 L^2(1-\theta)}{2} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \mathbb{E}_{\eta_m^{k,H^k-1}} \|g^{k,H^k-1}\|^2 \\
& \quad + \gamma\theta \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^k-1-1}} \mathbb{E}_{\eta_m^{k-1,H^k-1-1}} \|g^{k-1,H^k-1}\|^2 \\
& \quad + \frac{8\gamma^2 L(1-\theta)\delta_2}{\min_{1 \leq m \leq M} q_m} + \frac{4\gamma^2 L(1-\theta)\sigma^2}{\min_{1 \leq m \leq M} q_m}.
\end{aligned}$$

1191 We take expectation with respect to H^{k-1} and H^k , and apply Lemma 2:

$$\begin{aligned}
& \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \mathbb{E}_{\eta_m^{k,H^k-1}} [f(\tilde{x}^{k,H^k})] \\
& \leq (1-p) \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \mathbb{E}_{\eta_m^{k,H^k-1}} [f(\tilde{x}^{k,H^k})] \\
& \quad + p \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^k-1-1}} \mathbb{E}_{\eta_m^{k-1,H^k-1-1}} [f(\tilde{x}^{k,0})] \\
& \quad - \frac{\gamma(1-\theta)p}{4} \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^k-1-1}} \mathbb{E}_{\eta_m^{k-1,H^k-1-1}} \|\nabla f(x^{k,0})\|^2 \\
& \quad - \frac{\gamma(1-\theta)(1-p)}{4} \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \mathbb{E}_{\eta_m^{k,H^k-1}} \|\nabla f(x^{k,H^k})\|^2 \\
& \quad + \frac{\gamma^3 L^2(1-\theta)p}{2} \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^k-1-1}} \mathbb{E}_{\eta_m^{k-1,H^k-1-1}} \underbrace{\|g^{k,0}\|^2}_{=0} \\
& \quad + \frac{\gamma^3 L^2(1-\theta)(1-p)}{2} \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \mathbb{E}_{\eta_m^{k,H^k-1}} \|g^{k,H^k}\|^2 \\
& \quad + \gamma\theta \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^k-1-1}} \mathbb{E}_{\eta_m^{k-1,H^k-1-1}} \|g^{k-1,H^k-1}\|^2 \\
& \quad + \frac{8\gamma^2 L(1-\theta)\delta_2}{\min_{1 \leq m \leq M} q_m} + \frac{4\gamma^2 L(1-\theta)\sigma^2}{\min_{1 \leq m \leq M} q_m}. \tag{39}
\end{aligned}$$

1192 We use Lemma 4 to estimate $\|g^{k,H^k}\|^2$ and $\|g^{k-1,H^k-1}\|^2$. We have

$$\begin{aligned} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k,0}} \mathbb{E}_{\eta_m^{k,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \mathbb{E}_{\eta_m^{k,H^k-1}} \|g^{k,H^k}\|^2 &\leq \frac{96(1-\theta)^2 \alpha (\delta_1 + 1)}{p^2 \min_{1 \leq m \leq M} q_m} \mathbb{E}_{H^k} \|\nabla f(x^{k,H^k})\|^2 \\ &+ \frac{192(1-\theta)^2 \alpha \delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{96(1-\theta)^2 \alpha \sigma^2}{p^2 \min_{1 \leq m \leq M} q_m}. \end{aligned} \quad (40)$$

1193 Next, analogously to (19), we choose $\gamma \leq \frac{p \sqrt{\min_{1 \leq m \leq M} q_m}}{384L \sqrt{\alpha} \sqrt{\delta_1 + 1}}$ and obtain

$$\begin{aligned} &\mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \mathbb{E}_{\eta_m^{k-1,H^{k-1}-1}} \|g^{k-1,H^{k-1}}\|^2 \\ &\leq \frac{384(1-\theta)^2 \alpha (\delta_1 + 1)}{p^2 \min_{1 \leq m \leq M} q_m} \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \mathbb{E}_{\eta_m^{k-1,H^{k-1}-1}} \|\nabla f(x^{k,0})\|^2 \\ &+ \frac{384(1-\theta)^2 \alpha \delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{192(1-\theta)^2 \alpha \sigma^2}{p^2 \min_{1 \leq m \leq M} q_m}. \end{aligned} \quad (41)$$

1194 We combine (40) and (41) with (39) and use that H^{k-1} with $\{\xi_m^{k-1,h}\}_{h=0}^{H^{k-1}-1}$ and H^k with
 1195 $\{\eta_m^{k-1,h}\}_{h=0}^{H^{k-1}-1}$ are independent stochastic values. Moreover we take full expectation:

$$\begin{aligned} p \mathbb{E} [f(\tilde{x}^{k,H^k})] &\leq p \mathbb{E} [f(\tilde{x}^{k,0})] \\ &- \frac{\gamma(1-\theta)p}{4} \mathbb{E} \|\nabla f(x^{k,0})\|^2 - \frac{\gamma(1-\theta)(1-p)}{4} \mathbb{E} \|\nabla f(x^{k,H^k})\|^2 \\ &+ \frac{48\gamma^3 L^2 (1-\theta)^3 (1-p) \alpha (\delta_1 + 1)}{p^2 \min_{1 \leq m \leq M} q_m} \mathbb{E} \|\nabla f(x^{k,H^k})\|^2 \\ &+ \frac{192\gamma\theta(1-\theta)^2 \alpha (\delta_1 + 1)}{p^2 \min_{1 \leq m \leq M} q_m} \mathbb{E} \|\nabla f(x^{k,0})\|^2 \\ &+ \frac{96\gamma^3 L^2 (1-\theta)^3 (1-p) \alpha \delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{384\gamma\theta(1-\theta)^2 \alpha \delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{8\gamma^2 L (1-\theta) \delta_2}{\min_{1 \leq m \leq M} q_m} \\ &+ \frac{48\gamma^3 L^2 (1-\theta)^3 (1-p) \alpha \sigma^2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{192\gamma\theta(1-\theta)^2 \alpha \sigma^2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{4\gamma^2 L (1-\theta) \sigma^2}{\min_{1 \leq m \leq M} q_m} \\ &= p \mathbb{E} [f(\tilde{x}^{k,0})] \\ &- \frac{\gamma(1-\theta)(1-p)}{4} \left(1 - \frac{192\gamma^2 L^2 (1-\theta)^2 \alpha (\delta_1 + 1)}{p^2 \min_{1 \leq m \leq M} q_m} \right) \mathbb{E} \|\nabla f(x^{k,H^k})\|^2 \\ &- \frac{\gamma(1-\theta)p}{4} \left(1 - \frac{384\theta(1-\theta)^2 \alpha (\delta_1 + 1)}{p^3 \min_{1 \leq m \leq M} q_m} \right) \mathbb{E} \|\nabla f(x^{k,0})\|^2 \\ &+ \frac{96\gamma^3 L^2 (1-\theta)^3 (1-p) \alpha \delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{768\gamma\theta(1-\theta)^2 \alpha \delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{8\gamma^2 L (1-\theta) \delta_2}{\min_{1 \leq m \leq M} q_m} \\ &+ \frac{48\gamma^3 L^2 (1-\theta)^3 (1-p) \alpha \sigma^2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{192\gamma\theta(1-\theta)^2 \alpha \sigma^2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{4\gamma^2 L (1-\theta) \sigma^2}{\min_{1 \leq m \leq M} q_m}. \end{aligned}$$

1196 We choose $\theta \leq \frac{\gamma L p^2}{2} \gamma \leq \frac{p \min_{1 \leq m \leq M} q_m}{768L \alpha (\delta_1 + 1)}$. In that way,

$$\begin{aligned}
p\mathbb{E} \left[f(\tilde{x}^{k,H^k}) \right] &\leq p\mathbb{E} \left[f(\tilde{x}^{k,0}) \right] - \frac{\gamma(1-\theta)p}{8} \mathbb{E} \left\| \nabla f(x^{k,0}) \right\|^2 \\
&\quad + \frac{96\gamma^3 L^2 \alpha \delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{192\gamma^2 L \alpha \delta_2}{\min_{1 \leq m \leq M} q_m} + \frac{8\gamma^2 L \delta_2}{\min_{1 \leq m \leq M} q_m} \\
&\quad + \frac{48\gamma^3 L^2 \alpha \sigma^2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{96\gamma^2 L \alpha \sigma^2}{\min_{1 \leq m \leq M} q_m} + \frac{4\gamma^2 L \sigma^2}{\min_{1 \leq m \leq M} q_m}, \\
\frac{\gamma(1-\theta)}{8} \mathbb{E} \left\| \nabla f(x^{k,0}) \right\|^2 &\leq \mathbb{E} \left[f(\tilde{x}^{k,0}) \right] - \mathbb{E} \left[f(\tilde{x}^{k,H^k}) \right] \\
&\quad + \frac{96\gamma^3 L^2 \alpha \delta_2}{p^3 \min_{1 \leq m \leq M} q_m} + \frac{200\gamma^2 L \alpha \delta_2}{p \min_{1 \leq m \leq M} q_m} \\
&\quad + \frac{48\gamma^3 L^2 \alpha \sigma^2}{p^3 \min_{1 \leq m \leq M} q_m} + \frac{100\gamma^2 L \alpha \sigma^2}{p \min_{1 \leq m \leq M} q_m}.
\end{aligned}$$

1197 Note that $\tilde{x}^{k,H^k} = x^{k,H^k} - \gamma g^{k,H^k} = x^{k+1,0}$ and $\tilde{x}^{k,0} = x^{k,0}$. Thus,

$$\begin{aligned}
\frac{\gamma(1-\theta)}{8} \mathbb{E} \left\| \nabla f(x^{k,0}) \right\|^2 &\leq \mathbb{E} \left[f(x^{k,0}) \right] - \mathbb{E} \left[f(x^{k+1,0}) \right] + \frac{96\gamma^3 L^2 \alpha \delta_2}{p^3 \min_{1 \leq m \leq M} q_m} + \frac{200\gamma^2 L \alpha \delta_2}{p \min_{1 \leq m \leq M} q_m} \\
&\quad + \frac{48\gamma^3 L^2 \alpha \sigma^2}{p^3 \min_{1 \leq m \leq M} q_m} + \frac{100\gamma^2 L \alpha \sigma^2}{p \min_{1 \leq m \leq M} q_m}.
\end{aligned}$$

1198 Summing over all iterations, we obtain the result of the theorem:

$$\begin{aligned}
\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f(x^{k,0}) \right\|^2 &\leq \frac{8(f(x^{0,0}) - \mathbb{E} [f(x^{K,0})])}{\gamma(1-\theta)K} \\
&\quad + \frac{768\gamma^2 L^2 \alpha \delta_2}{p^3 \min_{1 \leq m \leq M} q_m (1-\theta)} + \frac{1600\gamma L \alpha \delta_2}{p \min_{1 \leq m \leq M} q_m (1-\theta)} \\
&\quad + \frac{384\gamma^2 L^2 \alpha \sigma^2}{p^3 \min_{1 \leq m \leq M} q_m (1-\theta)} + \frac{800\gamma L \alpha \sigma^2}{p \min_{1 \leq m \leq M} q_m (1-\theta)} \\
&\leq \frac{16(f(x^{0,0}) - f(x^*))}{\gamma K} \\
&\quad + \frac{1536\gamma^2 L^2 \alpha \delta_2}{p^3 \min_{1 \leq m \leq M} q_m} + \frac{3200\gamma L \alpha \delta_2}{p \min_{1 \leq m \leq M} q_m} \\
&\quad + \frac{768\gamma^2 L^2 \alpha \sigma^2}{p^3 \min_{1 \leq m \leq M} q_m} + \frac{1600\gamma L \alpha \sigma^2}{p \min_{1 \leq m \leq M} q_m}.
\end{aligned}$$

1199

□

1200 **Corollary 9 (Corollary 3).** Under conditions of Theorem 5 Algorithm 2 with fixed rules $\hat{\mathcal{R}} \equiv \tilde{\mathcal{R}} \equiv \mathcal{R}$
1201 needs

$$\begin{aligned}
&\mathcal{O} \left(\frac{M}{C} \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{\Delta L \alpha \delta_1}{\varepsilon^2} + \frac{\Delta L \alpha \delta_2}{\varepsilon^4} + \frac{\Delta L \alpha \sigma^2}{\varepsilon^4} \right) \right) \text{ epochs and} \\
&\mathcal{O} \left(M \frac{M}{C} \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{\Delta L \alpha \delta_1}{\varepsilon^2} + \frac{\Delta L \alpha \delta_2}{\varepsilon^4} + \frac{\Delta L \alpha \sigma^2}{\varepsilon^4} \right) \right) \text{ number of devices communications}
\end{aligned}$$

1202 to reach ε -accuracy, where $\varepsilon^2 = \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(x^{k,0})\|^2$, $\Delta = f(x^{0,0}) - f(x^*)$ and C is the number
 1203 of devices participating in each epoch.

1204 *Proof.* Proof is analogous to the proof of Corollary 5. □

1205 **Corollary 10.** Under conditions of Theorem 5 Algorithm 2 needs

$$\begin{aligned} & \mathcal{O} \left(\frac{M}{\min_{k,h} C^{k,h}} \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{\Delta L \alpha \delta_1}{\varepsilon^2} + \frac{\Delta L \alpha \delta_2}{\varepsilon^4} + \frac{\Delta L \alpha \sigma^2}{\varepsilon^4} \right) \right) \text{ epochs and} \\ & \mathcal{O} \left(M \left(\frac{M}{\min_{k,h} C^{k,h}} \right)^2 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{\Delta L \alpha \delta_1}{\varepsilon^2} + \frac{\Delta L \alpha \delta_2}{\varepsilon^4} + \frac{\Delta L \alpha \sigma^2}{\varepsilon^4} \right) \right) \end{aligned}$$

1206 number of devices communications to reach ε -accuracy, where $\varepsilon^2 = \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(x^{k,0})\|^2$, $\Delta =$
 1207 $f(x^{0,0}) - f(x^*)$ and $C^{k,h}$ is the number of devices participating in k -th iteration in h -th epoch.

1208 *Proof.* Proof is analogous to the proof of Corollary 6. □

1209 **Remark 3.** Considering fixed rules $\hat{\mathcal{R}} \equiv \tilde{\mathcal{R}} \equiv \mathcal{R}$,

1210 we have $\mathcal{O} \left(M \frac{M}{C} \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{\Delta L \delta_1}{\varepsilon^2} + \frac{\Delta L \delta_2}{\varepsilon^4} + \frac{\Delta L \sigma^2}{M \varepsilon^4} \right) \right)$

1211 and $\mathcal{O} \left(M^2 \frac{M}{C} \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{\Delta L \delta_1}{\varepsilon^2} + \frac{\Delta L \delta_2}{\varepsilon^4} + \frac{\Delta L \sigma^2}{M \varepsilon^4} \right) \right)$ number of devices communications with reg-
 1212 ularizing parameter $\alpha = 1$ and $\alpha = M$ respectively. Considering various rules, best case with

1213 regularizing coefficient $\alpha = 1$ gives us $\mathcal{O} \left(M \left(\frac{M}{\min_{k,h} C^{k,h}} \right)^2 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{\Delta L \delta_1}{\varepsilon^2} + \frac{\Delta L \delta_2}{\varepsilon^4} + \frac{\Delta L \sigma^2}{M \varepsilon^4} \right) \right)$

1214 and worst case $\alpha = M$ gives us $\mathcal{O} \left(M^2 \left(\frac{M}{\max_{k,h} C^{k,h}} \right)^2 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{\Delta L \delta_1}{\varepsilon^2} + \frac{\Delta L \delta_2}{\varepsilon^4} + \frac{\Delta L \sigma^2}{M \varepsilon^4} \right) \right)$ num-
 1215 ber of devices communications.

1216 E.2 Proof for strongly-convex setting

1217 **Theorem 6.** Suppose Assumptions 1, 2(b), 3, 4 hold. Then for Algorithm 2 with $\theta \leq \frac{p\gamma\mu}{4}$ and

1218 $\gamma \leq \frac{p^2 \min_{1 \leq m \leq M} q_m}{384 L \alpha (\delta_1 + 1)}$ it implies that

$$\mathbb{E} \|x^{K,0} - x^*\|^2 \leq \left(1 - \frac{\gamma\mu}{8}\right)^K \mathbb{E} \|x^{0,0} - x^*\|^2 + \frac{2368\gamma\alpha}{\mu p^3 \min_{1 \leq m \leq M} q_m} (2\delta_2 + \sigma^2).$$

1219 *Proof.* We start with the definition of virtual sequence:

$$\tilde{x}^{k,h} = x^{k,h} - \gamma \sum_{m=1}^M g_m^{k,h} = x^{k,h} - \gamma g^{k,h}. \quad (42)$$

1220 It is followed by

$$\begin{aligned} \tilde{x}^{k,h+1} &= x^{k,h+1} - \gamma \sum_{m=1}^M g_m^{k,h+1} \\ &= x^{k,h} - \gamma \left[(1 - \theta) \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \hat{\pi}_m^{k,h} \nabla f_m(x^{k,h}, \zeta_m^{k,h}) + \theta g^{k-1,H^{k-1}} \right] \\ &\quad - \gamma \sum_{m=1}^M g_m^{k,h} - \gamma (1 - \theta) \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \left(\frac{1}{M} - \hat{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \zeta_m^{k,h}) \end{aligned}$$

$$= \tilde{x}^{k,h} - \gamma \left[(1-\theta) \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) + \theta g^{k-1, H^{k-1}} \right]. \quad (43)$$

1221 Next, we use this to write a descent:

$$\begin{aligned} \|\tilde{x}^{k,h+1} - x^*\|^2 &= \|\tilde{x}^{k,h} - x^*\|^2 + 2 \langle \tilde{x}^{k,h} - x^*, \tilde{x}^{k,h+1} - \tilde{x}^{k,h} \rangle + \|\tilde{x}^{k,h+1} - \tilde{x}^{k,h}\|^2 \\ &\stackrel{(43)}{=} \|\tilde{x}^{k,h} - x^*\|^2 - 2\gamma\theta \langle \tilde{x}^{k,h} - x^*, g^{k-1, H^{k-1}} \rangle \\ &\quad - 2\gamma(1-\theta) \left\langle \tilde{x}^{k,h} - x^*, \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ &\quad + \gamma^2 \left\| \theta g^{k-1, H^{k-1}} + (1-\theta) \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\ &\stackrel{(\text{Jen})}{\leq} \|\tilde{x}^{k,h} - x^*\|^2 - 2\gamma\theta \langle \tilde{x}^{k,h} - x^*, g^{k-1, H^{k-1}} \rangle \\ &\quad - 2\gamma(1-\theta) \left\langle \tilde{x}^{k,h} - x^{k,h}, \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ &\quad - 2\gamma(1-\theta) \left\langle x^{k,h} - x^*, \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ &\quad + \gamma^2 \theta \|g^{k-1, H^{k-1}}\|^2 + \gamma^2 (1-\theta) \left\| \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2. \end{aligned}$$

1222 Now we use that $\eta_m^{k,h} \sim \mathcal{B}(q_m)$. Consequently, $\mathbb{E}\eta_m^{k,h} = q_m$. Since $\eta_m^{k,h}$ is independent of
1223 $x^{k,h}, \tilde{x}^{k,h}, \xi_m^{k,h}, g^{k-1, H^{k-1}}$, we take the expectation and obtain

$$\begin{aligned} \mathbb{E}_{\eta_m^{k,h}} \|\tilde{x}^{k,h+1} - x^*\|^2 &\leq \|\tilde{x}^{k,h} - x^*\|^2 - 2\gamma\theta \langle \tilde{x}^{k,h} - x^*, g^{k-1, H^{k-1}} \rangle \\ &\quad - 2\gamma(1-\theta) \left\langle \tilde{x}^{k,h} - x^{k,h}, \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ &\quad - 2\gamma(1-\theta) \left\langle x^{k,h} - x^*, \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ &\quad + \gamma^2 \theta \|g^{k-1, H^{k-1}}\|^2 \\ &\quad + \gamma^2 (1-\theta) \mathbb{E}_{\eta_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2. \end{aligned}$$

1224 Now we take the expectation over $\xi_m^{k,h}$. Mention that

$$\begin{aligned} \tilde{x}^{k,h} &\stackrel{(42)}{=} x^{k,h} - \gamma g^{k,h} \\ &\stackrel{\text{Line 12}}{=} x^{k,h} - \gamma \left(g^{k,h-1} + (1-\theta) \sum_{m=1}^M \frac{\eta_m^{k,h-1}}{q_m} \left(\frac{1}{M} - \hat{\pi}_m^{k,h-1} \right) \nabla f(x^{k,h-1}, \xi_m^{k,h-1}) \right). \end{aligned}$$

1225 Thus, $\tilde{x}^{k,h}$ and $\xi_m^{k,h}$ are independent. Analogously, $g^{k-1, H^{k-1}}$ and $\xi_m^{k,h}$ are independent. In this way,

$$\begin{aligned} \mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} \|\tilde{x}^{k,h+1} - x^*\|^2 &\leq \|\tilde{x}^{k,h} - x^*\|^2 - 2\gamma\theta \langle \tilde{x}^{k,h} - x^*, g^{k-1, H^{k-1}} \rangle \\ &\quad - 2\gamma(1-\theta) \langle \tilde{x}^{k,h} - x^{k,h}, \nabla f(x^{k,h}) \rangle \\ &\quad - 2\gamma(1-\theta) \langle x^{k,h} - x^*, \nabla f(x^{k,h}) \rangle \end{aligned}$$

$$\begin{aligned}
& + \gamma^2 \theta \left\| g^{k-1, H^{k-1}} \right\|^2 \\
& + \gamma^2 (1 - \theta) \mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2. \quad (44)
\end{aligned}$$

1226 Recall we estimated the last term in Theorem 5 in (37):

$$\begin{aligned}
\mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 & \leq \frac{16(\delta_1 + 1)}{\min_{1 \leq m \leq M} q_m} \left\| \nabla f(x^{k,h}) \right\|^2 + \frac{16\delta_2}{\min_{1 \leq m \leq M} q_m} \\
& + \frac{8\sigma^2}{\min_{1 \leq m \leq M} q_m}. \quad (45)
\end{aligned}$$

1227 Now let us estimate scalar products separately.

$$\begin{aligned}
-2\gamma\theta \left\langle \tilde{x}^{k,h} - x^*, g^{k-1, H^{k-1}} \right\rangle & \stackrel{(\text{Fen})}{\leq} \theta \left\| \tilde{x}^{k,h} - x^* \right\|^2 + \gamma^2 \theta \left\| g^{k-1, H^{k-1}} \right\|^2, \\
-2\gamma(1 - \theta) \left\langle \tilde{x}^{k,h} - x^{k,h}, \nabla f(x^{k,h}) \right\rangle & \stackrel{(\text{Fen})}{\leq} (1 - \theta) \left\| \tilde{x}^{k,h} - x^{k,h} \right\|^2 + \gamma^2 (1 - \theta) \left\| \nabla f(x^{k,h}) \right\|^2 \\
& \stackrel{(20)}{=} \gamma^2 (1 - \theta) \left\| g^{k,h} \right\|^2 + \gamma^2 (1 - \theta) \left\| \nabla f(x^{k,h}) \right\|^2, \\
-2\gamma(1 - \theta) \left\langle x^{k,h} - x^*, \nabla f(x^{k,h}) \right\rangle & \stackrel{\text{As. 2(b)}}{\leq} -\gamma\mu(1 - \theta) \left\| x^{k,h} - x^* \right\|^2 \\
& - 2\gamma(1 - \theta) [f(x^{k,h}) - f(x^*)] \\
& \stackrel{(\text{CS})}{\leq} -\frac{\gamma\mu(1 - \theta)}{2} \left\| \tilde{x}^{k,h} - x^* \right\|^2 \\
& + \gamma\mu(1 - \theta) \left\| x^{k,h} - \tilde{x}^{k,h} \right\|^2 \\
& - 2\gamma(1 - \theta) [f(x^{k,h}) - f(x^*)] \\
& \stackrel{(20)}{=} -\frac{\gamma\mu(1 - \theta)}{2} \left\| \tilde{x}^{k,h} - x^* \right\|^2 \\
& + \gamma^3\mu(1 - \theta) \left\| g^{k,h} \right\|^2 \\
& - 2\gamma(1 - \theta) [f(x^{k,h}) - f(x^*)].
\end{aligned}$$

1228 Substituting this estimates and (45) into (44),

$$\begin{aligned}
\mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} \left\| \tilde{x}^{k,h+1} - x^* \right\|^2 & \leq \left(1 - \frac{\gamma\mu(1 - \theta)}{2} + \theta \right) \left\| \tilde{x}^{k,h} - x^* \right\|^2 + 2\gamma^2\theta \left\| g^{k-1, H^{k-1}} \right\|^2 \\
& + \gamma^2(1 - \theta)(1 + \gamma\mu) \left\| g^{k,h} \right\|^2 \\
& + \frac{19\gamma^2(1 - \theta)(\delta_1 + 1)}{\min_{1 \leq m \leq M} q_m} \left\| \nabla f(x^{k,h}) \right\|^2 \\
& - 2\gamma(1 - \theta) [f(x^{k,h}) - f(x^*)] \\
& + \frac{16\gamma^2(1 - \theta)\delta_2}{\min_{1 \leq m \leq M} q_m} + \frac{8\gamma^2(1 - \theta)\sigma^2}{\min_{1 \leq m \leq M} q_m}.
\end{aligned}$$

1229 Let us choose $\theta \leq \frac{\gamma\mu}{4}$ and $\gamma \leq \frac{1}{L}$. Then, $\left(1 - \frac{\gamma\mu(1 - \theta)}{2} + \theta \right) \leq \left(1 - \frac{3\gamma\mu}{8} + \frac{\gamma\mu}{4} \right) = \left(1 - \frac{\gamma\mu}{8} \right)$. In
1230 this way,

$$\mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} \left\| \tilde{x}^{k,h+1} - x^* \right\|^2 \leq \left(1 - \frac{\gamma\mu}{8} \right) \left\| \tilde{x}^{k,h} - x^* \right\|^2 + 2\gamma^2\theta \left\| g^{k-1, H^{k-1}} \right\|^2$$

$$\begin{aligned}
& +\gamma^2(1-\theta)(1+\gamma\mu)\|g^{k,h}\|^2 \\
& +\frac{19\gamma^2(1-\theta)(\delta_1+1)}{\min_{1\leq m\leq M} q_m}\|\nabla f(x^{k,h})\|^2 \\
& -2\gamma(1-\theta)[f(x^{k,h})-f(x^*)] \\
& +\frac{16\gamma^2(1-\theta)\delta_2}{\min_{1\leq m\leq M} q_m}+\frac{8\gamma^2(1-\theta)\sigma^2}{\min_{1\leq m\leq M} q_m}.
\end{aligned}$$

1231 Next, we estimate

$$\frac{19\gamma^2(1-\theta)(\delta_1+1)}{\min_{1\leq m\leq M} q_m}\|\nabla f(x^{k,h})\|^2 \leq \frac{38\gamma^2 L(1-\theta)(\delta_1+1)}{\min_{1\leq m\leq M} q_m}[f(x^{k,h})-f(x^*)].$$

1232 It implies that

$$\begin{aligned}
\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\|\tilde{x}^{k,h+1}-x^*\|^2 & \leq \left(1-\frac{\gamma\mu}{8}\right)\|\tilde{x}^{k,h}-x^*\|^2+2\gamma^2\theta\|g^{k-1,H^{k-1}}\|^2 \\
& +\gamma^2(1-\theta)(1+\gamma\mu)\|g^{k,h}\|^2 \\
& -2\gamma(1-\theta)\left(1-\frac{19\gamma L(\delta_1+1)}{\min_{1\leq m\leq M} q_m}\right)[f(x^{k,h})-f(x^*)] \\
& +\frac{16\gamma^2(1-\theta)\delta_2}{\min_{1\leq m\leq M} q_m}+\frac{8\gamma^2(1-\theta)\sigma^2}{\min_{1\leq m\leq M} q_m}.
\end{aligned}$$

1233 Choosing $\gamma \leq \frac{\min_{1\leq m\leq M} q_m}{38L(\delta_1+1)}$, we can simplify as

$$\begin{aligned}
\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\|\tilde{x}^{k,h+1}-x^*\|^2 & \leq \left(1-\frac{\gamma\mu}{8}\right)\|\tilde{x}^{k,h}-x^*\|^2+2\gamma^2\theta\|g^{k-1,H^{k-1}}\|^2 \\
& +2\gamma^2(1-\theta)\|g^{k,h}\|^2 \\
& -\gamma(1-\theta)[f(x^{k,h})-f(x^*)] \\
& +\frac{16\gamma^2(1-\theta)\delta_2}{\min_{1\leq m\leq M} q_m}+\frac{8\gamma^2(1-\theta)\sigma^2}{\min_{1\leq m\leq M} q_m}.
\end{aligned}$$

1234 Now we put $h = H^k - 1$ and take additional expectations to obtain

$$\begin{aligned}
& \mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\dots\mathbb{E}_{\xi_m^{k,H^k-1}}\mathbb{E}_{\eta_m^{k,H^k-1}}\|\tilde{x}^{k,H^k}-x^*\|^2 \\
& \leq \left(1-\frac{\gamma\mu}{8}\right)\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\dots\mathbb{E}_{\xi_m^{k,H^k-1}}\mathbb{E}_{\eta_m^{k,H^k-1}}\|\tilde{x}^{k,H^k-1}-x^*\|^2 \\
& +2\gamma^2\theta\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\dots\mathbb{E}_{\xi_m^{k-1,H^k-1-1}}\mathbb{E}_{\eta_m^{k-1,H^k-1-1}}\|g^{k-1,H^{k-1}}\|^2 \\
& +2\gamma^2(1-\theta)\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\dots\mathbb{E}_{\xi_m^{k,H^k-1}}\mathbb{E}_{\eta_m^{k,H^k-1}}\|g^{k,H^k-1}\|^2 \\
& -\gamma(1-\theta)\mathbb{E}_{\xi_m^{k-1,0}}\mathbb{E}_{\eta_m^{k-1,0}}\dots\mathbb{E}_{\xi_m^{k,H^k-1}}\mathbb{E}_{\eta_m^{k,H^k-1}}[f(x^{k,H^k-1})-f(x^*)] \\
& +\frac{16\gamma^2(1-\theta)\delta_2}{\min_{1\leq m\leq M} q_m}+\frac{8\gamma^2(1-\theta)\sigma^2}{\min_{1\leq m\leq M} q_m}.
\end{aligned}$$

1235 We take expectation with respect to H^{k-1} and H^k , and apply Lemma 2:

$$\begin{aligned}
& \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^{k-1}-1}} \mathbb{E}_{\eta_m^{k,H^{k-1}-1}} \left\| \tilde{x}^{k,H^k} - x^* \right\|^2 \\
& \leq p \left(1 - \frac{\gamma\mu}{8} \right) \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \mathbb{E}_{\eta_m^{k-1,H^{k-1}-1}} \left\| \tilde{x}^{k,0} - x^* \right\|^2 \\
& \quad + (1-p) \left(1 - \frac{\gamma\mu}{8} \right) \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^{k-1}-1}} \mathbb{E}_{\eta_m^{k,H^{k-1}-1}} \left\| \tilde{x}^{k,H^k} - x^* \right\|^2 \\
& \quad + 2\gamma^2 \theta \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \mathbb{E}_{\eta_m^{k-1,H^{k-1}-1}} \left\| g^{k-1,H^{k-1}} \right\|^2 \\
& \quad + 2\gamma^2 (1-\theta)(1-p) \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^{k-1}-1}} \mathbb{E}_{\eta_m^{k,H^{k-1}-1}} \left\| g^{k,H^k} \right\|^2 \\
& \quad + 2\gamma^2 (1-\theta)p \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \mathbb{E}_{\eta_m^{k-1,H^{k-1}-1}} \underbrace{\left\| g^{k,0} \right\|^2}_{=0} \\
& \quad - \gamma(1-p)(1-\theta) \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^{k-1}-1}} \mathbb{E}_{\eta_m^{k,H^{k-1}-1}} \left[f(x^{k,H^k}) - f(x^*) \right] \\
& \quad - \gamma p(1-\theta) \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \mathbb{E}_{\eta_m^{k-1,H^{k-1}-1}} \left[f(x^{k,0}) - f(x^*) \right] \\
& \quad + \frac{16\gamma^2(1-\theta)\delta_2}{\min_{1 \leq m \leq M} q_m} + \frac{8\gamma^2(1-\theta)\sigma^2}{\min_{1 \leq m \leq M} q_m}. \tag{46}
\end{aligned}$$

1236 We use Lemma 4 to estimate $\left\| g^{k,H^k} \right\|^2$ and $\left\| g^{k-1,H^{k-1}} \right\|^2$. We have

$$\begin{aligned}
\mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k,0}} \mathbb{E}_{\eta_m^{k,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \mathbb{E}_{\eta_m^{k,H^k-1}} \left\| g^{k,H^k} \right\|^2 & \leq \frac{96(1-\theta)^2\alpha(\delta_1+1)}{p^2 \min_{1 \leq m \leq M} q_m} \mathbb{E}_{H^k} \left\| \nabla f(x^{k,H^k}) \right\|^2 \\
& \quad + \frac{192(1-\theta)^2\alpha\delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{96(1-\theta)^2\alpha\sigma^2}{p^2 \min_{1 \leq m \leq M} q_m}. \tag{47}
\end{aligned}$$

1237 Next, analogously to (29), we choose $\gamma \leq \frac{p\sqrt{\min_{1 \leq m \leq M} q_m}}{384L\sqrt{\alpha}\sqrt{\delta_1+1}}$ and obtain

$$\begin{aligned}
& \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \mathbb{E}_{\eta_m^{k-1,H^{k-1}-1}} \left\| g^{k-1,H^{k-1}} \right\|^2 \\
& \leq \frac{384(1-\theta)^2\alpha(\delta_1+1)}{p^2 \min_{1 \leq m \leq M} q_m} \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \mathbb{E}_{\eta_m^{k-1,H^{k-1}-1}} \left\| \nabla f(x^{k,0}) \right\|^2 \\
& \quad + \frac{384(1-\theta)^2\alpha\delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{192(1-\theta)^2\alpha\sigma^2}{p^2 \min_{1 \leq m \leq M} q_m}. \tag{48}
\end{aligned}$$

1238 Now we use (Lip) and that H^k with $\{\xi_m^{k-1,h}\}_{h=0}^{H^{k-1}-1}$ and H^{k-1} with $\{\eta_m^{k-1,h}\}_{h=0}^{H^{k-1}-1}$ are independent stochastic values. Moreover, we combine (47) and (48) with (46) and take full expectation.

$$\begin{aligned}
p\mathbb{E} \left\| \tilde{x}^{k,H^k} - x^* \right\|^2 & \leq p \left(1 - \frac{\gamma\mu}{8} \right) \mathbb{E} \left\| \tilde{x}^{k,0} - x^* \right\|^2 \\
& \quad - \gamma(1-p)(1-\theta) \mathbb{E} \left[f(x^{k,H^k}) - f(x^*) \right] \\
& \quad - \gamma p(1-\theta) \mathbb{E} \left[f(x^{k,0}) - f(x^*) \right] \\
& \quad + \frac{384\gamma^2 L(1-\theta)^3(1-p)\alpha(\delta_1+1)}{p^2 \min_{1 \leq m \leq M} q_m} \mathbb{E} \left[f(x^{k,H^k}) - f(x^*) \right] \\
& \quad + \frac{1536\gamma^2 L\theta(1-\theta)^2\alpha(\delta_1+1)}{p^2 \min_{1 \leq m \leq M} q_m} \mathbb{E} \left[f(x^{k,0}) - f(x^*) \right]
\end{aligned}$$

$$\begin{aligned}
& + \frac{384\gamma^2(1-\theta)^3(1-p)\alpha\delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{768\gamma^2\theta(1-\theta)^2\alpha\delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{16\gamma^2(1-\theta)\delta_2}{\min_{1 \leq m \leq M} q_m} \\
& + \frac{192\gamma^2(1-\theta)^3(1-p)\alpha\sigma^2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{384\gamma^2\theta(1-\theta)^2\alpha\sigma^2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{8\gamma^2(1-\theta)\sigma^2}{\min_{1 \leq m \leq M} q_m} \\
\leq & p \left(1 - \frac{\gamma\mu}{8}\right) \mathbb{E} \|\tilde{x}^{k,0} - x^*\|^2 \\
& - \gamma(1-p)(1-\theta) \left(1 - \frac{384\gamma L(1-\theta)^2\alpha(\delta_1+1)}{p^2 \min_{1 \leq m \leq M} q_m}\right) \\
& \cdot \mathbb{E} [f(x^{k,H^k}) - f(x^*)] \\
& - \gamma p(1-\theta) \left(1 - \frac{1536\gamma L\theta(1-\theta)\alpha(\delta_1+1)}{p^3 \min_{1 \leq m \leq M} q_m}\right) \mathbb{E} [f(x^{k,0}) - f(x^*)] \\
& + \frac{384\gamma^2(1-\theta)^3(1-p)\alpha\delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{768\gamma^2\theta(1-\theta)^2\alpha\delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{16\gamma^2(1-\theta)\delta_2}{\min_{1 \leq m \leq M} q_m} \\
& + \frac{192\gamma^2(1-\theta)^3(1-p)\alpha\sigma^2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{384\gamma^2\theta(1-\theta)^2\alpha\sigma^2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{8\gamma^2(1-\theta)\sigma^2}{\min_{1 \leq m \leq M} q_m}.
\end{aligned}$$

1240 Choosing $\theta \leq \frac{p\gamma\mu}{4}$ and $\gamma \leq \frac{p^2 \min_{1 \leq m \leq M} q_m}{384L\alpha(\delta_1+1)}$, we obtain

$$\mathbb{E} \|\tilde{x}^{k,H^k} - x^*\|^2 \leq \left(1 - \frac{\gamma\mu}{8}\right) \mathbb{E} \|\tilde{x}^{k,0} - x^*\|^2 + \frac{296\gamma\alpha}{p^3 \min_{1 \leq m \leq M} q_m} (2\delta_2 + \sigma^2).$$

1241 Note that $\tilde{x}^{k,H^k} = x^{k,H^k} - \gamma g^{k,H^k} = x^{k+1,0}$ and $\tilde{x}^{k,0} = x^{k,0}$. Thus,

$$\mathbb{E} \|x^{k+1,0} - x^*\|^2 \leq \left(1 - \frac{\gamma\mu}{8}\right) \mathbb{E} \|x^{k,0} - x^*\|^2 + \frac{296\gamma^2\alpha}{p^3 \min_{1 \leq m \leq M} q_m} (2\delta_2 + \sigma^2).$$

1242 It remains for us to going into recursion over all epochs and the result of the theorem:

$$\begin{aligned}
\mathbb{E} \|x^{K,0} - x^*\|^2 & \leq \left(1 - \frac{\gamma\mu}{8}\right)^K \mathbb{E} \|x^{0,0} - x^*\|^2 + \frac{296\gamma^2\alpha}{p^3 \min_{1 \leq m \leq M} q_m} (2\delta_2 + \sigma^2) \sum_{k=0}^K \left(1 - \frac{\gamma\mu}{8}\right)^k \\
& \leq \left(1 - \frac{\gamma\mu}{8}\right)^K \mathbb{E} \|x^{0,0} - x^*\|^2 + \frac{2368\gamma\alpha}{\mu p^3 \min_{1 \leq m \leq M} q_m} (2\delta_2 + \sigma^2).
\end{aligned}$$

1243

□

1244 **Corollary 11 (Corollary 4).** Under conditions of Theorem 6 Algorithm 2 with fixed rules $\hat{\mathcal{R}} \equiv \tilde{\mathcal{R}} \equiv$
1245 \mathcal{R} needs

$$\begin{aligned}
& \tilde{\mathcal{O}} \left(\left(\frac{M}{C}\right)^2 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{L}{\mu} \alpha \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{C} \frac{\alpha \delta_2}{\mu^2 \varepsilon} + \frac{M}{C} \frac{\alpha \sigma^2}{\mu^2 \varepsilon} \right) \right) \\
& \text{epochs and} \\
& \tilde{\mathcal{O}} \left(M \left(\frac{M}{C}\right)^2 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{L}{\mu} \alpha \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{C} \frac{\alpha \delta_2}{\mu^2 \varepsilon} + \frac{M}{C} \frac{\alpha \sigma^2}{\mu^2 \varepsilon} \right) \right) \\
& \text{number of devices communications}
\end{aligned}$$

1246 to reach ε -accuracy, where $\varepsilon^2 = \mathbb{E} \|x^{K,0} - x^*\|^2$ and C is number of devices participating in each
 1247 epoch.

1248 *Proof.* Proof is analogous to the proof of Corollary 7. □

1249 **Corollary 12.** Under conditions of Theorem 6 Algorithm 2 needs

$$\tilde{\mathcal{O}} \left(\left(\frac{M}{\min_{k,h} C^{k,h}} \right)^2 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{L}{\mu} \alpha \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{\min_{k,h} C^{k,h}} \frac{\alpha \delta_2}{\mu^2 \varepsilon} + \frac{M}{\min_{k,h} C^{k,h}} \frac{\alpha \sigma^2}{\mu^2 \varepsilon} \right) \right)$$

epochs or

$$\tilde{\mathcal{O}} \left(M \left(\frac{M}{\min_{k,h} C^{k,h}} \right)^3 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{L}{\mu} \alpha \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{\min_{k,h} C^{k,h}} \frac{\alpha \delta_2}{\mu^2 \varepsilon} + \frac{M}{\min_{k,h} C^{k,h}} \frac{\alpha \sigma^2}{\mu^2 \varepsilon} \right) \right)$$

communications

1250 to reach ε -accuracy, where $\varepsilon^2 = \mathbb{E} \|x^{K,0} - x^*\|^2$ and $C^{k,h}$ is the number of devices participating in
 1251 k -th iteration in h -th epoch.

1252 *Proof.* Proof is analogous to the proof of Corollary 8. □

1253 **Remark 4.** Considering fixed rules $\hat{\mathcal{R}} \equiv \tilde{\mathcal{R}} \equiv \mathcal{R}$,

1254 we have $\tilde{\mathcal{O}} \left(M \left(\frac{M}{C} \right)^2 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{L}{\mu} \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{C} \frac{\delta_2}{\mu^2 \varepsilon} + \frac{M}{C} \frac{\sigma^2}{\mu^2 \varepsilon} \right) \right)$

1255 and $\tilde{\mathcal{O}} \left(M^2 \left(\frac{M}{C} \right)^2 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{L}{\mu} \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{C} \frac{\delta_2}{\mu^2 \varepsilon} + \frac{M}{C} \frac{\sigma^2}{\mu^2 \varepsilon} \right) \right)$ number of devices commu-
 1256 nications with regularizing parameter $\alpha = 1$ and $\alpha = M$ respectively. Con-
 1257 sidering various rules, best case with regularizing coefficient $\alpha = 1$ gives us

1258 $\tilde{\mathcal{O}} \left(M \left(\frac{M}{\min_{k,h} C^{k,h}} \right)^3 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{L}{\mu} \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{\min_{k,h} C^{k,h}} \frac{\delta_2}{\mu^2 \varepsilon} + \frac{M}{\min_{k,h} C^{k,h}} \frac{\sigma^2}{\mu^2 \varepsilon} \right) \right)$ and worst case $\alpha =$

1259 M gives us $\tilde{\mathcal{O}} \left(M^2 \left(\frac{M}{\min_{k,h} C^{k,h}} \right)^3 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{L}{\mu} \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{\min_{k,h} C^{k,h}} \frac{\delta_2}{\mu^2 \varepsilon} + \frac{M}{\min_{k,h} C^{k,h}} \frac{\sigma^2}{\mu^2 \varepsilon} \right) \right)$ num-
 1260 ber of devices communications.