

A STRATEGY-AGNOSTIC FRAMEWORK FOR PARTIAL PARTICIPATION IN FEDERATED LEARNING

000
001
002
003
004
005 **Anonymous authors**
006 Paper under double-blind review
007
008
009
010

ABSTRACT

011 Partial participation (PP) is a fundamental paradigm in federated learning, where
012 only a fraction of clients can be involved in each communication round. In recent
013 years, a wide range of mechanisms for partial participation have been proposed.
014 However, the effectiveness of a particular technique strongly depends on problem-
015 specific characteristics, e.g. local data distributions. Consequently, achieving better
016 performance requires a comprehensive search across a number of strategies. This
017 observation highlights the necessity of a unified framework. In this paper, we
018 address this challenge by introducing a general scheme that can be combined with
019 almost any client selection strategy. We provide a unified theoretical analysis
020 of our approach without relying on properties specific to individual heuristics.
021 Furthermore, we extend it to settings with unstable client-server connections,
022 thereby covering real-world scenarios in federated learning. We present empirical
023 validation of our framework across a range of PP strategies on image classification
024 tasks, employing modern architectures, such as FasterViT.
025

1 INTRODUCTION

026 Optimization is a cornerstone of training machine learning and neural network models. In a nutshell,
027 almost every AI-based solution aims to minimize an empirical risk (Shalev-Shwartz et al., 2010),
028 which evaluates how well the data is approximated. This process involves adjusting parameters
029 to reduce the discrepancy between predicted outputs and ground truth labels, thereby improving
030 generalization performance. Formally, the problem can be expressed as
031

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

032 where x denotes the trainable parameters of the model g , (a_i, b_i) is the i -th sample from the dataset
033 with size n , and ℓ is the loss function. Nowadays, there is a variety of methods developed to efficiently
034 solve equation 1 (Robbins and Monro, 1951; Nesterov, 1983; Kingma and Ba, 2014; Defazio and
035 Mishchenko, 2023). The current successes of machine/deep learning owe much to the development
036 of powerful numerical techniques that enable training on a huge amount of samples. Large-scale
037 data processing became possible with the advancement of distributed optimization (Verbraeken
038 et al., 2020). Instead of solving the problem on a single machine, samples are shared among M
039 nodes/devices/clients/machines connected via a server. Hence, the problem equation 1 transforms
040 into
041

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

042 where n_m is the size of the dataset, stored on m -th device.
043

1.1 CLIENT WEIGHTING

044 Parallel data processing helps to reduce computational time significantly (Zinkevich et al., 2010;
045 Abadi et al., 2016; Jouppi et al., 2017). However, contemporary applications present new challenges.
046 Training samples are often accumulated locally by each specific machine, rather than being collected
047 and distributed manually. This paradigm with data remaining on edge devices is called federated
048 learning (Konečný et al., 2016; McMahan et al., 2017; Bonawitz et al., 2019). In such a setup, local
049 datasets are typically heterogeneous – they vary in size, distribution, and quality. For instance, one
050 device may hold unique objects that are poorly represented across the rest of the network, but are
051
052
053

054 crucial for capturing more dependencies. This leads to the conclusion that some clients may be more
 055 useful than others. Modern approaches usually assign dynamic weights $\{\pi_m\}_{m=1}^M$ and use
 056

$$057 \quad f(x) = \sum_{m=1}^M \pi_m f_m(x), \text{ s.t. } \pi_m > 0, \sum_{m=1}^M \pi_m = 1 \quad (3)$$

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

062 **Weighting Based on Data Quality/Quantity.** The most straightforward way to cope with data
 063 imbalance is to consider a number of local samples. McMahan et al. (2017) suggested setting each
 064 coefficient as the constant $\pi_m = n_m/n$. Since then, many modifications of this approach have been
 065 proposed, including federated averaging schemes with momentum (Wang et al., 2019; Reddi et al.,
 066 2020), variance reduction (Liang et al., 2019; Karimireddy et al., 2020) and proximal updates (Li
 067 et al., 2020). However, this type of weighting ignores heterogeneity in terms of data quality, leading
 068 to bias, e.g. if some client holds an enormous amount of objects with the same labels. To support the
 069 diversity of training samples, Yurochkin et al. (2019) proposed to match the neurons of client neural
 070 networks before averaging. Building on the foundations laid by this work, subsequent works have
 071 explored more efficient approaches extensively (Wang et al., 2020a; Zhang et al., 2022; Yang et al.,
 072 2023; Wu et al., 2023; Kafshgari et al., 2023).

073 **Learned Weighting Strategies.** It is also common to learn weighting strategies instead of using
 074 fixed heuristics. Mohri et al. (2019) were among the first to present results in this direction. They pro-
 075 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
 076 to well-trained devices. The idea of optimizing agnostic empirical loss was then generalized by Li
 077 et al. (2019a). Their q-FedAvg can be reduced to agnostic optimization as one of the special cases.
 078 However, in practice, it is hard to search for appropriate saddle-points (Daskalakis and Panageas,
 079 2018; Jin et al., 2020), especially in federated learning (Sharma et al., 2023). As a result, the commu-
 080 nity has shifted towards softer adaptive approaches based on local losses (Zhang et al., 2020; Gao
 081 et al., 2022) and gradients (Wang et al., 2020b; Luo et al., 2024).

082 **Robust Weighting.** The idea of assigning weights to the devices found its application in robust
 083 optimization, where malicious clients can disrupt the learning process (Baruch et al., 2019; Xie et al.,
 084 2020; Fang et al., 2020). To combat such attacks, advanced schemes usually compute $\{\pi_m\}_{m=1}^M$, as
 085 the trust scores of the devices based on their objectives decrease (Xie et al., 2019), local gradients
 086 (Cao et al., 2020; Yan et al., 2023), and the number of local samples (Cao and Lai, 2019). Recently,
 087 researchers came up with the idea of using a Bayesian approach (Yang et al., 2024).

089 1.2 CLIENT SAMPLING

090 Another significant issue of federated learning, on par with heterogeneity, is the communication
 091 bottleneck (Tang et al., 2020; Shi et al., 2020). Sharing information between machines is costly and
 092 can limit the positive effect of parallelism, which is especially tangible when clients send messages
 093 to the server (Kairouz et al., 2021). This issue is magnified in federated learning, where edge devices
 094 may have unstable network connectivity, and transmitting large updates may be prohibitively slow.
 095 Many techniques exist to reduce communication (Seide et al., 2014; Alistarh et al., 2017; Stich,
 096 2018). Partial participation is a special one among them (Li et al., 2019b; Yang et al., 2021). In each
 097 communication round, only a random subset of clients participates in training, while the rest remain
 098 inactive. This approach offloads the server by decreasing the number of updates that need to be
 099 aggregated. Moreover, it provides significant advantages in edge computing, where communication
 100 channels are not equivalent, or some of them may be unavailable. Nowadays, there is a wide range of
 101 heuristics, which allows to choose subset of clients efficiently.

102 **Data-Based Sampling Strategies.** Methods from this class rely on zero- and first-order information
 103 of local functions. Importance Sampling FedAvg (Rizk et al., 2021) was one of the first
 104 such approaches. The authors suggested evaluating the relevance of a device by how large its gradient
 105 is relative to the others. Indeed, a small gradient makes a weak contribution to the step. Consequently,
 106 communication with this node can be neglected. Nguyen et al. (2020) proposed an orthogonal
 107 approach. Their FOLB measures the angle between local and average gradient. If it is negative, then
 108 such a device is useless at the current moment. This idea was then developed extensively in (Wu

108 and Wang, 2022; Zhou et al., 2022). In addition, techniques based on the norms of updates (Chen
 109 et al., 2020) and local loss decrease (Cho et al., 2022) were proposed. There are also a number of
 110 approaches that dynamically exploit data heterogeneity to maintain balance (Zhang et al., 2023) or
 111 support diversity (Chen and Vikalo, 2024).

112 **System-Based Sampling Strategies.** Another approach is to use information about the network
 113 itself. FedCS (Nishio and Yonetani, 2019) categorizes clients into groups based on their computa-
 114 tional power. This strategy saves wall-clock time by avoiding frequent selection of weak devices.
 115 Another class of techniques optimizes energy consumption (Xu and Wang, 2020). Most modern
 116 system heterogeneity techniques also incorporate local data considerations (Lai et al., 2021; Li et al.,
 117 2022). F3AST (Ribero et al., 2022) learns an availability-dependent client selection strategy to
 118 minimize the impact of variance on the global model’s convergence.

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

124 1.3 UNIFICATION OF SAMPLING STRATEGIES

125 Existing papers in this area of research are built around the federated averaging scheme (McMahan
 126 et al., 2017). Li et al. (2019b) proposed an analysis for strongly convex objectives, obtaining a
 127 sublinear convergence rate $\mathcal{O}(\kappa^2/K)$, where κ is the condition number. However, they modeled the
 128 partial participation environment via unbiased sampling. Cho et al. (2022) were the first to study the
 129 unified case with biased devices selection. They derived $\mathcal{O}(\kappa^2/K + \kappa Q)$, where Q is a non-vanishing
 130 term that becomes zero solely in the absence of sampling bias. Thus, the authors recovered the results
 131 of Li et al. (2019b), but failed to extend the theory to weaker assumptions. The first success in this
 132 direction was achieved in (Luo et al., 2022). This work resolved key questions regarding biased
 133 sampling in the strongly convex case. However, the non-convex analysis holds greater significance
 134 for applications. For this setting, Wang and Ji (2022) obtained $\mathcal{O}(\sqrt{L}/\sqrt{K} + \delta)$, where L is the
 135 smoothness constant and δ is the uniform bound on the difference between local gradients. This result
 136 contains the non-vanishing term and does not match the lower bound $\Omega(L/K)$ (Carmon et al., 2020).
 137 Thus, current works in this field rely on FedAvg. Consequently, their analysis requires boundedness
 138 of gradients (Li et al., 2019b; Cho et al., 2022; Luo et al., 2022) or their differences (Wang and Ji,
 139 2022) even in the non-stochastic case. Therefore, there is still no flawless unified theory of partial
 140 participation.

141 1.4 OUR CONTRIBUTION

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

150 2 SETUP

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

153 **Assumption 2.1.** The function f is L -smooth, i.e. for all $x, y \in \mathbb{R}^d$ it satisfies

$$154 \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

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

158 **Assumption 2.2.** The function f is:

162 (a) **non-convex** with at least one global minimum:
 163 there exists may be not unique, x^* s.t. $f(x^*) = \inf_{x \in \mathbb{R}^d} f(x) > -\infty$.
 164

165 (b) **μ -strongly convex**, i.e. for all $x, y \in \mathbb{R}^d$ it satisfies
 166
$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2.$$

168 Federated learning methods usually require a bound on data heterogeneity to provide convergence
 169 guarantees (Khaled et al., 2020; Karimireddy et al., 2020). In our work, we quantify it via gradients
 170 (Tang et al., 2018; Stich, 2020).

171 **Assumption 2.3.** Each gradient ∇f_m is similar to the full gradient ∇f , i.e. for all $x \in \mathbb{R}^d$ it satisfies

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

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

179 **Assumption 2.4.** Each worker has access to a stochastic gradient $\nabla f_m(x, \xi_m)$. This is an unbiased
 180 random variable with bounded variance, i.e. for all $x \in \mathbb{R}^d$ it satisfies

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

184 This assumption appears in different forms in a number of classic papers (Stich, 2018; Gower et al.,
 185 2019; Gorbunov et al., 2020). Next, we consider that weights $\{\pi_m\}_{m=1}^M$ from equation 3 lie on the
 186 regularized simplex. Namely, $\pi \in \Delta_1^M \cap \left(\bigcap_{m=1}^M \{\pi : e_m^\top \pi + \frac{\alpha}{M} \geq 0\} \right)$, where $1 \leq \alpha \leq M$ is the
 187 regularization parameter and e is the unit basis. This technique is useful for solving a wide range of
 188 tasks (Mehta et al., 2024).

190 3 ALGORITHMS AND ANALYSIS

192 3.1 MOTIVATION

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

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

208 Viewing partial participation as weight vector sparsification reveals connections to well-studied
 209 techniques (Beznosikov et al., 2023). A state-of-the-art approach to handle it efficiently is error
 210 feedback (Stich and Karimireddy, 2020; Richtárik et al., 2021). Since sampling rules are represented
 211 as compressors, we believe that this idea may be extremely useful in our setting as well. However,
 212 we cannot apply the error feedback framework directly. The reason is that the sampling rules are
 213 non-contractive compressors, as they zero out certain local gradients. Formally, there does not exist
 214 $\beta < \infty$ such that $\|x - \mathcal{C}(x)\|^2 \leq \left(1 - \frac{1}{\beta}\right) \|x\|^2$ for $\mathcal{C}(x) = 0, x \in \mathbb{R}^d$.

215 Thus, we have to address the challenge of designing a scheme that can handle non-contractive
 216 compression before proceeding to a unified analysis of partial participation.

216 3.2 PARTIAL PARTICIPATION WITHOUT UNAVAILABLE DEVICES
217

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

222 **Algorithm 1** PPBC

```

223 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  $M$ 
224 2: Parameters: Stepsize  $\gamma > 0$ , momentum  $0 < \theta < 1$ , regularization  $1 \leq \alpha \leq M$ 
225 3: for epochs  $k = 0, \dots, K - 1$  do
226 4:   Initialize  $\pi^k$  // Server weighs clients using any procedure
227 5:    $\hat{\pi}^k = \hat{\mathcal{R}}^k(\pi^k)$  // Server selects clients to communicate through epoch using any rule  $\hat{\mathcal{R}}$ 
228 6:    $g_m^{k,0} = 0$  // Each client initializes the gradient surrogate
229 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
230 8:   Generate  $H^k \sim \text{Geom}(p)$  // Server generates number of iterations of  $k$ -th epoch
231 9:   for iterations  $h = 0, \dots, H^k - 1$  do
232 10:     $\tilde{\pi}^{k,h} = \tilde{\mathcal{R}}^{k,h}(\hat{\pi}^k)$  // Server selects clients to communicate at the current round using rule  $\tilde{\mathcal{R}}$ 
233 11:    for devices  $m = 1 \dots M$  in parallel do
234 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
235 13:    end for
236 14:    for each device  $m : \tilde{\pi}_m^{k,h} \neq 0$  do
237 15:      Send  $\nabla f_m(x^{k,h}, \xi_m^{k,h})$  to the server
238 16:    end for
239 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
240 18:    parameters
241 19:  end for
242 20:  for devices  $m = 1 \dots M$  in parallel do
243 21:    Send  $g_m^{k, H^k}$  to the server
244 22:  end for
245 23:   $g^{k, H^k} = \sum_{m=1}^M g_m^{k, H^k}$  // Server aggregates gradient surrogates
246 24: end for

```

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

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

$$263 \tilde{x}^{k,h} = x^{k,h} - \gamma \sum_{m=1}^M g_m^{k,h}$$

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

$$265 \tilde{x}^{k,h+1} = \tilde{x}^{k,h} - \gamma \left[(1 - \theta) \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) + \theta g^{k-1, H^{k-1}} \right].$$

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

Lemma 3.1. *Suppose Assumptions 2.3, 2.4 hold. We consider the epoch size $H^k \sim \text{Geom}(p)$ and $1 \leq \alpha \leq M$. Then for Algorithm 1 it implies*

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

Assumption 2.4 is required only to handle local stochasticity. If the devices are able to compute exact gradients, Lemma 3.1 holds with $\sigma = 0$. For the details, see Appendix D. As a result, we obtain the convergence theorem.

Theorem 3.2. *Suppose Assumptions 2.1, 2.2(a), 2.3, 2.4 hold. Then for Algorithm 1 with $\theta \leq \frac{\gamma L p^2}{2}$ and $\gamma \leq \frac{p}{384L\alpha(\delta_1+1)}$ it implies that*

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

The main obstacle in proving Theorem 3.2 is the terms $\|g^{k, H^k}\|^2$ and $\|g^{k-1, H^{k-1}}\|^2$ that appear in the analysis. Using Lemma 3.1, they can be screwed to $\|\nabla f(x^{k, H^k})\|^2$ and $\|\nabla f(x^{k-1, H^{k-1}})\|^2$, respectively. The first norm is easy to analyze. Classically, it serves as a convergence criterion. Eliminating the second one turns out to be challenging. To cope with it, we incorporate the surrogate into the starting point of the epoch (Line 7). For the details, see Appendix D.1. With such an estimate, there is a technique to choose the stepsize γ appropriately to obtain convergence (Stich, 2019).

Corollary 3.3. *Under conditions of Theorem 3.2 Algorithm 1 with fixed rules $\hat{\mathcal{R}}^k \equiv \tilde{\mathcal{R}}^{k,h} \equiv \mathcal{R}$ needs*

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

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

We also consider varying sampling rules $\hat{\mathcal{R}}^k$ and $\tilde{\mathcal{R}}^{k,h}$ to study corollaries of Theorem 3.2.

Corollary 3.4. *Under conditions of Theorem 3.2 Algorithm 1 needs*

$$\mathcal{O}\left(\frac{M}{\min_{k,h} C^{k,h}} \left(\frac{\Delta L\alpha\delta_1}{\varepsilon^2} + \frac{\Delta L\alpha\delta_2}{\varepsilon^4} + \frac{\Delta L\alpha\sigma^2}{M\varepsilon^4} \right)\right) \text{ epochs and}$$

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

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 number of devices participating in k -th iteration in h -th epoch.

In our work, the analysis is extended to the strongly convex case.

324 **Theorem 3.5.** Suppose Assumptions 2.1, 2.2(b), 2.3, 2.4 hold. Then for Algorithm 1 with $\theta \leq \frac{p\gamma\mu}{4}$
 325 and $\gamma \leq \frac{p^2}{96L\alpha(\delta_1+1)}$ it implies that
 326

$$327 \mathbb{E} \|x^{K,0} - x^*\|^2 \leq \left(1 - \frac{\gamma\mu}{8}\right)^K \|x^{0,0} - x^*\|^2 + \frac{8\gamma\alpha}{\mu p^3} \left(144\delta_2 + \frac{74\sigma^2}{M}\right).$$

329 As well as for the non-convex objective, suitable γ can be chosen in Theorem 3.5.
 330

331 **Corollary 3.6.** Under conditions of Theorem 3.5 Algorithm 1 with fixed rules $\widehat{\mathcal{R}}^{k,h} \equiv \widetilde{\mathcal{R}}^{k,h} \equiv \mathcal{R}$
 332 needs

$$333 \widetilde{\mathcal{O}} \left(M \left(\frac{M}{C} \right)^2 \left(\frac{L}{\mu} \alpha \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{C} \frac{\alpha \delta_2}{\mu^2 \varepsilon} + \frac{\alpha \sigma^2}{\mu^2 C \varepsilon} \right) \right)$$

335 number of devices communications to reach ε -accuracy, where $\varepsilon^2 = \mathbb{E} \|x^{K,0} - x^*\|^2$ and C is the
 336 number of devices participating in each epoch.
 337

338 **Corollary 3.7.** Under conditions of Theorem 3.5 Algorithm 1 needs

$$339 \widetilde{\mathcal{O}} \left(\left(\frac{M}{\min_{k,h} C^{k,h}} \right)^2 \left(\frac{L}{\mu} \alpha \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{\min_{k,h} C^{k,h}} \frac{\alpha \delta_2}{\mu^2 \varepsilon} + \frac{\alpha \sigma^2}{\mu^2 \min_{k,h} C^{k,h} \varepsilon} \right) \right) \text{ epochs and}$$

$$343 \widetilde{\mathcal{O}} \left(M \left(\frac{M}{\min_{k,h} C^{k,h}} \right)^3 \left(\frac{L}{\mu} \alpha \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{\min_{k,h} C^{k,h}} \frac{\alpha \delta_2}{\mu^2 \varepsilon} + \frac{\alpha \sigma^2}{\mu^2 \min_{k,h} C^{k,h} \varepsilon} \right) \right)$$

347 number of devices communications to reach ε -accuracy, where $\varepsilon^2 = \mathbb{E} \|x^{K,0} - x^*\|^2$ and $C^{k,h}$ is
 348 the number of devices participating in k -th iteration in h -th epoch.
 349

350 3.3 PARTIAL PARTICIPATION WITH UNAVAILABLE DEVICES

351 The previous section addresses partial participation when all devices are available to communicate
 352 with the server. Indeed, in Algorithm 1 each node receives the current parameters at the end of the
 353 iteration, but does not send its gradient. This is motivated by the fact that forwarding a message
 354 from the client to the server is much more expensive than the other way around (Kairouz et al.,
 355 2021). However, in practice, some devices can become inactive periodically (Li et al., 2019b; Yang
 356 et al., 2021). Namely, these machines not only refrain from transmitting information but also do not
 357 perform local computations. In this section, we extend our theory to cover the case where the actual
 358 parameters are sent to only a fraction of the clients.
 359

360 **Description of Algorithm 2.** In this section we present the part of Algorithm 2 (see Appendix A)
 361 that reflects key differences from Algorithm 1. To design it, we refuse using the biased sampling
 362 rule $\widetilde{\mathcal{R}}$ during the epoch. Instead, we simulate outage probability of the m -th device as a Bernoulli
 363 random variable $\eta_m^{k,h} \sim \text{Be}(q_m)$ (Chung, 2000) (Line 11). To describe client disconnection formally,
 364 $\eta_m^{k,h}$ is used to update the gradient surrogates (Line 12) and to perform the step (Line 17). Thus, in
 365 practice, it is not necessary for an inactive device to know the actual parameters. We also normalize
 366 the computed gradients by factors $\{q_m\}_{m=1}^M$ to balance their magnitudes.
 367

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

368 12: $g_m^{k,h+1} = g_m^{k,h} + (1 - \theta) \frac{\eta_m^{k,h}}{q_m} \left(\frac{1}{M} - \hat{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h})$

369 17: $x^{k,h+1} = x^{k,h} - \gamma \left[(1 - \theta) \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \hat{\pi}_m^{k,h} \nabla f_m(x^{k,h}, \xi_m^{k,h}) + \theta g^{k-1, H^{k-1}} \right]$
 370

372 **Analysis of Algorithm 2.** We formulate the results for both non-convex and strongly-convex cases.
 373

374 **Corollary 3.8.** Suppose Assumptions 2.1, 2.2(a), 2.3, 2.4 hold. Algorithm 2 with fixed rules $\widehat{\mathcal{R}}^k \equiv$
 375 $\widetilde{\mathcal{R}}^{k,h} \equiv \mathcal{R}$ needs

$$376 \mathcal{O} \left(M \frac{M}{C} \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{\Delta L \alpha \delta_1}{\varepsilon^2} + \frac{\Delta L \alpha \delta_2}{\varepsilon^4} + \frac{\Delta L \alpha \sigma^2}{\varepsilon^4} \right) \right)$$

378 number of devices communications to reach ε -accuracy, where $\varepsilon^2 = \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(x^{k,0})\|^2$, $\Delta =$
 379 $f(x^{0,0}) - f(x^*)$ and C is the number of devices participating in each epoch.
 380
 381

382 **Corollary 3.9.** Under conditions of Theorem E.2 Algorithm 2 needs

$$383 \mathcal{O} \left(\frac{M}{\min_{k,h} C^{k,h}} \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{\Delta L \alpha \delta_1}{\varepsilon^2} + \frac{\Delta L \alpha \delta_2}{\varepsilon^4} + \frac{\Delta L \alpha \sigma^2}{\varepsilon^4} \right) \right) \text{ epochs and}$$

$$386 \mathcal{O} \left(M \left(\frac{M}{\min_{k,h} C^{k,h}} \right)^2 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{\Delta L \alpha \delta_1}{\varepsilon^2} + \frac{\Delta L \alpha \delta_2}{\varepsilon^4} + \frac{\Delta L \alpha \sigma^2}{\varepsilon^4} \right) \right)$$

390 number of devices communications to reach ε -accuracy, where $\varepsilon^2 = \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(x^{k,0})\|^2$, $\Delta =$
 391 $f(x^{0,0}) - f(x^*)$ and $C^{k,h}$ is the number of devices participating in k -th iteration in h -th epoch.
 392
 393

394 **Corollary 3.10.** Suppose Assumptions 2.1, 2.2(b), 2.3, 2.4 hold. Algorithm 2 with fixed rules
 395 $\widehat{\mathcal{R}}^k \equiv \widetilde{\mathcal{R}}^{k,h} \equiv \mathcal{R}$ needs

$$396 \mathcal{O} \left(M \left(\frac{M}{C} \right)^2 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{L}{\mu} \alpha \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{C} \frac{\alpha \delta_2}{\mu^2 \varepsilon} + \frac{M}{C} \frac{\alpha \sigma^2}{\mu^2 \varepsilon} \right) \right)$$

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

401 **Corollary 3.11.** Under conditions of Theorem E.6 Algorithm 2 needs

$$403 \widetilde{\mathcal{O}} \left(\left(\frac{M}{\min_{k,h} C^{k,h}} \right)^2 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{L}{\mu} \alpha \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{\min_{k,h} C^{k,h}} \frac{\alpha \delta_2}{\mu^2 \varepsilon} + \frac{M}{\min_{k,h} C^{k,h}} \frac{\alpha \sigma^2}{\mu^2 \varepsilon} \right) \right)$$

406 epochs or

$$408 \widetilde{\mathcal{O}} \left(M \left(\frac{M}{\min_{k,h} C^{k,h}} \right)^3 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{L}{\mu} \alpha \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{\min_{k,h} C^{k,h}} \frac{\alpha \delta_2}{\mu^2 \varepsilon} + \frac{M}{\min_{k,h} C^{k,h}} \frac{\alpha \sigma^2}{\mu^2 \varepsilon} \right) \right)$$

411 communications

412 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
 413 k -th iteration in h -th epoch.
 414

415 For more details, see Appendix E. Note that $\min_{1 \leq m \leq M} q_m$ is a constant lying in the interval $(0, 1]$.
 416 Thus, the rates of Algorithm 2 do not differ significantly from those for Algorithm 1. The only
 417 deterioration occurs in the variance term associated with local stochasticity. Thus, if each device has
 418 an access to its exact gradient, there is no asymptotical difference compared to Corollaries 3.3 and
 419 3.6.

420 3.4 DISCUSSION

421 We analyzed a wide class of sampling and weighting techniques and proposed algorithms for different
 422 network scenarios. Their rates asymptotically coincide with the optimal ones for SGD-like approaches
 423 (Stich, 2019). Due to considering biased strategies, we obtained an additional factor M/C . Again
 424 analogizing to compression, this multiplier signifies compression power. It is a well-known fact
 425 that there is no theoretical improvement for methods built upon error-feedback (Richtárik et al.,
 426 2021; Beznosikov et al., 2023). However, we recover the convergence of SGD in the case of full
 427 participation. Comparing our non-convex rate regarding the main term $\mathcal{O}(1/\varepsilon^2)$ with prior works,
 428 we note that it surpasses that in (Wang and Ji, 2022) ($\mathcal{O}(1/\varepsilon^4 + \delta_2)$) both asymptotically and by
 429 the absence of the non-vanishing term. Next, comparing strongly-convex rates ($\mathcal{O}(\kappa \log 1/\varepsilon)$), we
 430 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,
 431 both of these works lack non-convex analysis. We highlight that we soften assumptions from all
 432 aforementioned works.

432 4 EXPERIMENTS

434 To validate our theoretical findings, we conduct a systematic empirical comparison of six optimization
 435 frameworks – FedAvg (Reddi et al., 2020), SCAFFOLD (Karimireddy et al., 2020), FedDyn
 436 (Chen et al., 2023), Moon (Li et al., 2021), and PPBC (Algorithm 1) — evaluated under full client
 437 participation (FCP), along with two additional frameworks – F3AST (Ribero et al., 2022) and
 438 PPBC+ (Algorithm 2) — specifically designed for and evaluated under partial client participation
 439 (PCP). Crucially, we fix the sampling strategy across all frameworks to isolate how each optimizer
 440 interacts with it, thereby decoupling the sampling mechanism from core algorithmic innovations for
 441 FCP experiments. All methods are compared under identical experimental conditions: same model
 442 architectures, benchmark datasets, and hardware configurations. The following section details the
 443 experimental setup, including architectures, datasets, and infrastructure.

444 **Experimental Setup.** We evaluate sampling strategies under three distinct data distribution settings:
 445 **(distr-1)** homogeneous (i.i.d.), **(distr-2)** heterogeneous (client-specific class sets), and **(distr-3)**
 446 strongly heterogeneous (varying data volumes and class skew). In this section we will present results
 447 for the most challenging setup with **distr-3**, full version of experiments is in Appendix B along with
 448 other details. Experiments use CIFAR-10 (Krizhevsky et al., 2009) with RESNET-18 (Meng et al.,
 449 2019) for image classification and FOOD101 Bossard et al. (2014) with FASTERViT (Hatamizadeh
 450 et al., 2023) for fine-tuning, providing a controlled benchmark for comparing Algorithm 1. Import-
 451 antly, each plot compares frameworks – not strategies – by fixing the underlying strategy and varying
 452 the framework. This correspondence is formalized in Algorithm 1, where the gradient surrogate
 453 term vanishes, recovering the conventional update rule. Further implementation details (partitioning,
 454 architecture, datasets) appear in Appendix B.

455 4.1 FULL CLIENT PARTICIPATION

456 **Client Selection Rule.** Notably, not all strategies in-
 457 cluded in our comparative analysis inherently incorporate
 458 a client selection mechanism. To ensure a fair and con-
 459 sistent evaluation, we uniformly applied the following
 460 selection rule across all methods:

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

461 where Top_C denotes taking $C > 0$ clients with the high-
 462 est weights π^k . Consequently, the remainder of our ex-
 463 periments will focus exclusively on the formulation and
 464 analysis of weight update rules, while treating the client
 465 selection process itself as a fixed component of the exper-
 466 imental framework.

467 **Client Sampling.** We evaluate four established client
 468 sampling strategies, each designed to improve convergence
 469 or robustness by prioritizing clients based on different cri-
 470 teria. PoC (Cho et al., 2022) selects clients proportionally
 471 to their local loss values, favoring those with higher em-
 472 pirical risk to accelerate optimization. BANT (Xie et al., 2019)
 473 employs a trust-based mechanism, dynamically scoring
 474 clients by their historical alignment with server-side val-
 475 idation performance, thereby promoting reliability over
 476 time. FOLB (Nguyen et al., 2020) samples clients based on the projected utility of their updates
 477 – specifically, the inner product between local gradients and the server’s global descent direction –
 478 to maximize progress per round. Finally, GNS (Wang et al., 2020b) prioritizes clients with larger
 479 gradient norms, under the intuition that clients exhibiting stronger local signals contribute more
 480 meaningfully to global updates.

481 Full algorithmic descriptions and implementation details for all strategies are provided in Appendix B.

482 **Results.** The comparative results are summarized in Table 1, with primary evaluation
 483 based on final test loss and accuracy metric. Figure 1 complements this by visu-
 484 alizing the training dynamics of our PPBC framework against the strongest baselines.
 485 For FedAvg and SCAFFOLD, we report their best-performing variant per sampling strat-

Table 1: Frameworks and strategies comparison on CIFAR-10 & RESNET-18.

Method + Strategy	distr-3	
	Loss (↓)	Acc (↑)
FedAvg + PoC	0.898±0.021	65.3±0.20
FedAvg + FOLB	0.674±0.020	71.42±0.19
FedAvg + BANT	2.324±0.023	11.32±0.25
FedAvg + GNS	0.657±0.019	71.15±0.19
SCAFFOLD + PoC	0.788±0.020	69.81±0.19
SCAFFOLD + FOLB	0.663±0.016	71.80±0.20
SCAFFOLD + BANT	0.698±0.017	71.31±0.18
SCAFFOLD + GNS	0.689±0.020	71.75±0.19
FedDyn	0.652±0.016	76.71±0.14
Moon	0.627±0.014	75.21±0.15
PPBC + PoC	0.367±0.019	88.87±0.16
PPBC + FOLB	0.362±0.016	88.91±0.14
PPBC + BANT	0.357±0.015	88.96±0.15
PPBC + GNS	0.364±0.016	88.90±0.15

Notation: All values averaged over 3 seeds. Arrows indicate optimization direction: ↓ minimize loss, ↑ maximize accuracy. Green color represents our algorithms.

486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
egy, ensuring a fair and strategy-aware comparison. This allows us to isolate the impact of the optimization framework itself, independent of sampling-induced variance.

4.2 PARTIAL CLIENT PARTICIPATION

Client Sampling and Partial Participation. To simulate real-world scenarios, we model client presence at each round via independent Bernoulli trials with participation probability q_m . We evaluate performance across a spectrum of participation regimes, ranging from full availability ($q_m = 1$) to highly sparse communication ($q_m = 0.3$), reflecting scenarios with frequent dropouts or intermittent connectivity. To contextualize our framework’s robustness under such conditions, we include comparative experiments against F3AST, an algorithm specifically designed to handle client outages and non-uniform participation.

For PPBC+, we set server strategy $\widehat{\mathcal{R}}^k$ with the FOLB strategy and employ PoC as the client sampling mechanism $\widetilde{\mathcal{R}}^k$.

Results. Similarly to the previous section, results are summarized in Table 2, with the primary evaluation based on the final test loss and accuracy metrics. Figure 2 represents accuracy graphs of our PPBC+ framework (Algorithm 2) with $q_m = 0.3$ against F3AST with $q_m = 1, 0.7, 0.5$ and FedAvg with $q_m = 1$. This plot clearly demonstrates the superiority of our method over F3AST. Moreover, we highlight that even under the most challenging communication conditions ($q_m = 0.3$), our approach consistently converges to substantially higher accuracy than all competing baselines.

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 learning scenarios. Results demonstrate a substantial performance gap between conventional approaches (FedAvg, SCAFFOLD, FedDyn, Moon) and Algorithm 1. Additionally, we analyzed the behavior of the PPBC+ (Algorithm 2) 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 3, 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.

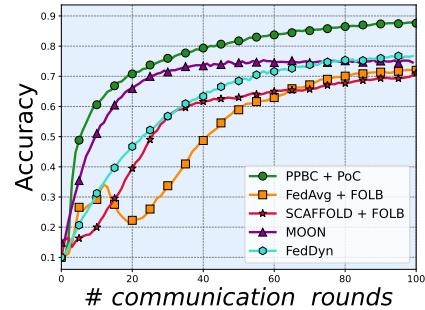


Figure 1: Comparison graphs on **distr-3** for best runs.

Table 2: Frameworks and strategies comparison on FASTERViT & FOOD101.

Method	distr-3	
	Loss (↓)	Acc (↑)
FedAvg ($q_m = 1$)	1.896±0.021	56.74±0.13
F3AST ($q_m = 1$)	1.692±0.022	68.31±0.11
F3AST ($q_m = 0.7$)	1.754±0.020	65.52±0.12
F3AST ($q_m = 0.5$)	1.812±0.018	61.30±0.13
PPBC+ ($q_m = 1$)	0.930±0.017	76.11±0.09
PPBC+ ($q_m = 0.7$)	0.937±0.018	76.04±0.12
PPBC+ ($q_m = 0.5$)	0.961±0.018	75.07±0.10
PPBC+ ($q_m = 0.3$)	0.996±0.020	74.68±0.11

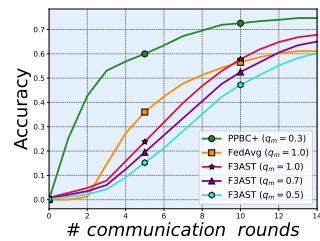


Figure 2: Comparison graphs on **distr-3** for best runs.

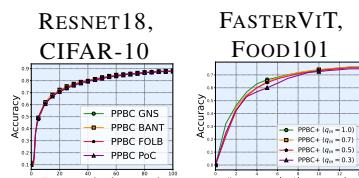


Figure 3: Test accuracy of PPBC/PPBC+ for image classification with RESNET18 on CIFAR-10 and FASTERViT fine-tuning on FOOD101.

540 REFERENCES
541

542 Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu
543 Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. {TensorFlow}: a system for
544 {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and
545 implementation (OSDI 16)*, pages 265–283, 2016.

546 Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-
547 efficient sgd via gradient quantization and encoding. *Advances in neural information processing
548 systems*, 30, 2017.

549 Zeyuan Allen-Zhu. Katyusha x: Practical momentum method for stochastic sum-of-nonconvex
550 optimization. *arXiv preprint arXiv:1802.03866*, 2018.

551 Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for
552 distributed learning. *Advances in Neural Information Processing Systems*, 32, 2019.

553 Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression
554 for distributed learning. *Journal of Machine Learning Research*, 24(276):1–50, 2023.

555 Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir
556 Ivanov, Chloe Kiddon, Jakub Konečný, Stefano Mazzocchi, Brendan McMahan, et al. Towards
557 federated learning at scale: System design. *Proceedings of machine learning and systems*, 1:
558 374–388, 2019.

559 Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative compo-
560 nents with random forests. In *European Conference on Computer Vision*, 2014.

561 Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. Fltrust: Byzantine-robust federated
562 learning via trust bootstrapping. *arXiv preprint arXiv:2012.13995*, 2020.

563 Xinyang Cao and Lifeng Lai. Distributed gradient descent algorithm robust to an arbitrary number of
564 byzantine attackers. *IEEE Transactions on Signal Processing*, 67(22):5850–5864, 2019.

565 Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary
566 points i. *Mathematical Programming*, 184(1):71–120, 2020. doi: 10.1007/s10107-019-01406-y.
567 URL <https://doi.org/10.1007/s10107-019-01406-y>.

568 Huancheng Chen and Haris Vikalo. Heterogeneity-guided client sampling: Towards fast and efficient
569 non-iid federated learning. *Advances in Neural Information Processing Systems*, 37:65525–65561,
570 2024.

571 Shuaijun Chen, Omid Tavallaie, Michael Henri Hambali, Seid Miad Zandavi, Hamed Haddadi,
572 Nicholas Lane, Song Guo, and Albert Y Zomaya. Optimization of federated learning’s client
573 selection for non-iid data based on grey relational analysis. *arXiv preprint arXiv:2310.08147*,
574 2023.

575 Wenlin Chen, Samuel Horvath, and Peter Richtarik. Optimal client sampling for federated learning.
576 *arXiv preprint arXiv:2010.13723*, 2020.

577 Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Towards understanding biased client selection in
578 federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages
579 10351–10375. PMLR, 2022.

580 Kai Lai Chung. *A course in probability theory*. Elsevier, 2000.

581 George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control,
582 signals and systems*, 2(4):303–314, 1989.

583 Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in
584 min-max optimization. *Advances in neural information processing systems*, 31, 2018.

585 Aaron Defazio and Konstantin Mishchenko. Learning-rate-free learning by d-adaptation. In *Inter-
586 national Conference on Machine Learning*, pages 7449–7479. PMLR, 2023.

594 Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. Local model poisoning attacks to
 595 $\{\text{Byzantine-Robust}\}$ federated learning. In *29th USENIX security symposium (USENIX Security*
 596 *20*), pages 1605–1622, 2020.

597

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

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

604

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

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

611

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

615

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

620

621 Zahra Hafezi Kafshgari, Chamani Shiranthika, Parvaneh Saeedi, and Ivan V Bajić. Quality-adaptive
 622 split-federated learning for segmenting medical images with inaccurate annotations. In *2023 IEEE*
623 20th International Symposium on Biomedical Imaging (ISBI), pages 1–5. IEEE, 2023.

624

625 Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin
 626 Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*,
 627 14(1–2):1–210, 2021.

628

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

631

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

634

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

637

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

641

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

643

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

646

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

648 Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of*
 649 *the IEEE/CVF conference on computer vision and pattern recognition*, pages 10713–10722, 2021.
 650

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

653 Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith.
 654 Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*,
 655 2:429–450, 2020.

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

658 Xianfeng Liang, Shuheng Shen, Jingchang Liu, Zhen Pan, Enhong Chen, and Yifei Cheng. Variance
 659 reduced local sgd with lower communication complexity. *arXiv preprint arXiv:1912.12844*, 2019.
 660

661 Bing Luo, Wenli Xiao, Shiqiang Wang, Jianwei Huang, and Leandros Tassiulas. Tackling system and
 662 statistical heterogeneity for federated learning with adaptive client sampling. In *IEEE INFOCOM
 663 2022-IEEE conference on computer communications*, pages 1739–1748. IEEE, 2022.

664

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

667

668 Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas.
 669 Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence
 670 and statistics*, pages 1273–1282. PMLR, 2017.

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

674

675 Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. Frame attention networks for facial expression
 676 recognition in videos. In *2019 IEEE international conference on image processing (ICIP)*, pages
 677 3866–3870. IEEE, 2019.

678 Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International
 679 conference on machine learning*, pages 4615–4625. PMLR, 2019.

680

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

682

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

686

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

689

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

693

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

697

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

700

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

702 Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. Ef21: A new, simpler, theoretically better,
 703 and practically faster error feedback. *Advances in Neural Information Processing Systems*, 34:
 704 4384–4396, 2021.

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

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

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

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

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

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

723 Shaohuai Shi, Zhenheng Tang, Xiaowen Chu, Chengjian Liu, Wei Wang, and Bo Li. A quantitative
 724 survey of communication optimizations in distributed deep learning. *IEEE Network*, 35(3):230–237,
 725 2020.

726 Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*,
 727 2018.

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

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

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

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

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

742 Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S
 743 Rellermeyer. A survey on distributed machine learning. *Acm computing surveys (csur)*, 53(2):
 744 1–33, 2020.

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

748 Jianyu Wang, Vinayak Tantia, Nicolas Ballas, and Michael Rabbat. Slowmo: Improving
 749 communication-efficient distributed sgd with slow momentum. *arXiv preprint arXiv:1910.00643*,
 750 2019.

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

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

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

759

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

762

763 Cong Xie, Sanmi Koyejo, and Indranil Gupta. Zeno: Distributed stochastic gradient descent with
 764 suspicion-based fault-tolerance. In *International Conference on Machine Learning*, pages 6893–
 765 6901. PMLR, 2019.

766

767 Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Fall of empires: Breaking byzantine-tolerant sgd
 768 by inner product manipulation. In *Uncertainty in Artificial Intelligence*, pages 261–270. PMLR,
 769 2020.

770

771 Jie Xu and Heqiang Wang. Client selection and bandwidth allocation in wireless federated learning
 772 networks: A long-term perspective. *IEEE Transactions on Wireless Communications*, 20(2):
 773 1188–1200, 2020.

774

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

778

779 Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation
 780 in non-iid federated learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:
 781 5974–5986, 2021. NeurIPS 2021.

782

783 Mingkun Yang, Ran Zhu, Qing Wang, and Jie Yang. Fedtrans: Client-transparent utility estimation for
 784 robust federated learning. In *The Twelfth International Conference on Learning Representations*,
 785 2024.

786

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

790

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

794

795 Jianyi Zhang, Ang Li, Minxue Tang, Jingwei Sun, Xiang Chen, Fan Zhang, Changyou Chen, Yiran
 796 Chen, and Hai Li. Fed-cbs: A heterogeneity-aware client sampling mechanism for federated
 797 learning via class-imbalance reduction. In *International Conference on Machine Learning*, pages
 798 41354–41381. PMLR, 2023.

799

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

803

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

806

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

810

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

813

814

815

816

817

818

819

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

810	APPENDIX	
811		
812	CONTENTS	
813		
814	1 Introduction	1
815	1.1 Client Weighting	1
816	1.2 Client Sampling	2
817	1.3 Unification of Sampling Strategies	3
818	1.4 Our Contribution	3
819		
820	2 Setup	3
821		
822	3 Algorithms and Analysis	4
823	3.1 Motivation	4
824	3.2 Partial Participation without Unavailable Devices	5
825	3.3 Partial Participation with Unavailable Devices	7
826	3.4 Discussion	8
827		
828	4 Experiments	9
829	4.1 Full client participation	9
830	4.2 Partial client participation	10
831		
832	A Partial Participation with Unavailable Devices	17
833		
834	B Additional experiments and details	17
835		
836	C General statements	22
837		
838	D Proofs for Algorithm 1	22
839	D.1 Proof for non-convex case	26
840	D.2 Proof for strongly-convex case	33
841		
842	E Proofs for Algorithm 2	40
843	E.1 Proof for non-convex setting	43
844	E.2 Proof for strongly-convex setting	49
845		
846		
847		
848		
849		
850		
851		
852		
853		
854		
855		
856		
857		
858		
859		
860		
861		
862		
863		

864 **A PARTIAL PARTICIPATION WITH UNAVAILABLE DEVICES**
865866 In this section, we present Algorithm 2, which is the complete version of algorithm from Section
867 3.3. This method can be applied to environments where devices do not perform local computations
868 periodically.869 **Algorithm 2** PPBC+
870

```

871 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  $M$ 
872 2: Parameters: Stepsize  $\gamma > 0$ , momentum  $0 < \theta < 1$ , regularization  $1 \leq \alpha \leq M$ 
873 3: for epochs  $k = 0, \dots, K - 1$  do
874 4:   Initialize  $\pi^k$  // Server weighs clients using any procedure
875 5:    $\hat{\pi}^k = \hat{\mathcal{R}}^k(\pi^k)$  // Server selects clients to communicate through epoch using any rule  $\hat{\mathcal{R}}$ 
876 6:    $g_m^{k,0} = 0$  // Each client initializes the gradient surrogate
877 7:    $x^{k,0} = x^{-1, H^{k-1}} - \gamma g^{-1, H^{k-1}}$  // Server initializes the initial point of the epoch
878 8:   Generate  $H^k \sim \text{Geom}(p)$  // Server generates number of iterations of  $k$ -th epoch
879 9:   for iterations  $h = 0, \dots, H^k - 1$  do
880 10:    for devices  $m = 1 \dots M$  in parallel do
881 11:      Generate  $\eta_m^{k,h} \sim \mathcal{B}(q_m)$  // Device generates its state: available / unavailable
882 12:       $g_m^{k,h+1} = g_m^{k,h} + (1 - \theta) \frac{\eta_m^{k,h}}{q_m} \left( \frac{1}{M} - \hat{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h})$  // Update the gradient
883      surrogate
884 13:    end for
885 14:    for each device  $m : \eta_m^{k,h} \neq 0$  and  $\hat{\pi}_m^{k,h} \neq 0$  do
886 15:      Send  $\frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h})$  to the server
887 16:    end for
888 17:     $x^{k,h+1} = x^{k,h} - \gamma \left[ (1 - \theta) \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \hat{\pi}_m^{k,h} \nabla f_m(x^{k,h}, \xi_m^{k,h}) + \theta g^{k-1, H^{k-1}} \right]$  // Server up-
889      dates parameters
890 18:  end for
891 19:  for devices  $m = 1 \dots M$  in parallel do
892 20:    Send  $g_m^{k, H^k}$  to the server
893 21:  end for
894 22:   $g^{k, H^k} = \sum_{m=1}^M g_m^{k, H^k}$  // Server aggregates gradient surrogates
895 23: end for

```

896 **B ADDITIONAL EXPERIMENTS AND DETAILS**
897898 Our code is available at https://anonymous.4open.science/r/EF25_ICLR/.
899900 **Hardware Details.** The experiments were conducted using Python with the PyTorch deep learning
901 framework (Paszke et al., 2017). The computational hardware consisted of a server equipped with
902 an Intel Xeon Gold 6342 CPU and two NVIDIA A100 40GB GPUs. The total runtime for all
903 experimental evaluations amounted to approximately 80 hours. To simulate a federated learning
904 environment, data was distributed across clients based on a heterogeneity parameter.
905906 **Data Distribution.** In our study, we employed 10 clients for both the RESNET-18 on CIFAR-10
907 setup and the FASTERViT fine-tuning on the FOOD101 dataset. This client count was carefully
908 chosen to enable comprehensive evaluation across the diverse data distribution scenarios proposed
909 in our work, while maintaining computational feasibility for thorough experimentation. Below, we
910 provide a detailed summary of the data distribution characteristics for each experimental setup.
911912 Homogeneous data distribution (**distr-1**) – each client has the same number of data samples, and
913 class labels are uniformly distributed across clients.
914915 *Example (CIFAR-10):* Each client has 500 training samples per class, resulting in 5,000 samples per
916 client in total.
917Heterogeneous data distribution (**distr-2**) – each client has the same total number of samples, but
918 class labels are distributed in a non-IID manner.
919

918 *Example (CIFAR-10):* We split the 10 classes into two disjoint groups (e.g., classes 0-4 and 5-9), and
 919 assign clients to one of the two groups. Clients in each group receive data only from their assigned
 920 classes. Additionally, the number of samples per class varies across clients.

921 Pathological data distribution (**distr-3**) – clients possess
 922 different amounts of data. The distribution of sample
 923 proportions across clients is as follows:

924 Within each client, class labels are sampled according
 925 to a Dirichlet distribution with concentration parameter
 926 $\alpha = 0.5$, resulting in highly non-IID label distributions.

927 Next, we provide a detailed overview of the client
 928 sampling strategies and present comparative results for
 929 FedAvg, SCAFFOLD, and Algorithm 1. We exclude
 930 FedDyn and Moon from this analysis, as their designs
 931 incorporate fixed strategies that cannot be decoupled from
 932 their core update rules.

933 **Loss-aware Client Sampling.** Building upon previous
 934 work, Cho et al. (2022) introduced the POWER-OF-
 935 CHOICE (P_OC) strategy, which employs a weighted client sampling mechanism based on local loss
 936 values. Formally, the weight update rule can be expressed as:

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

$$938 \quad p_m = \frac{n_m}{\left(\sum_{m'=1}^M n_{m'} \right)}.$$

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

$$942 \quad \pi^k = \left(\left[\frac{1}{n_m} \sum_{i_m=1}^{n_m} \ell(g(x, a_{i_m}), b_{i_m}) \right] \right)_{m=1}^M.$$

943 **Trust-Score Sampling.** The study by Xie et al. (2019) introduces the BANT, which implements
 944 a trust-based sampling mechanism. This approach assigns dynamic trust scores to clients based on
 945 historical performance metrics. Thus, weight update rule can be described as:

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

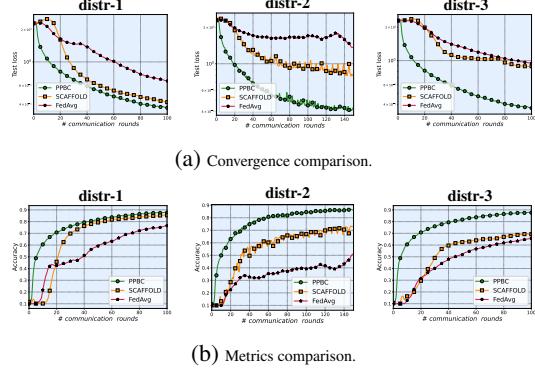
$$949 \quad TS_m^k = \exp \left[-\frac{1}{|\mathcal{V}|} \sum_{\xi \in \mathcal{V}} f_m(x^k, \xi) \right].$$

950 2. The weights are updated with a probability
 951 proportional to trust scores:

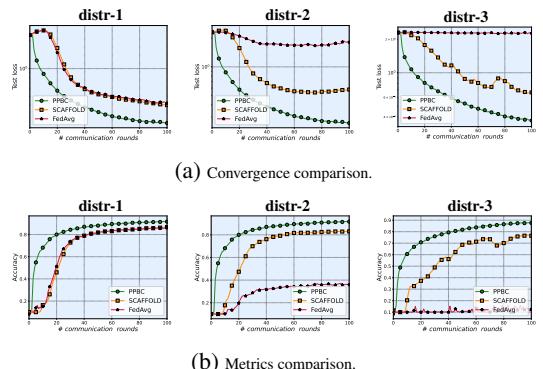
$$952 \quad \pi^k = \left(\frac{TS_m^k}{\sum_{m'=1}^M TS_{m'}^k} \right)_{m=1}^M.$$

953 Table 3: Client-wise data sample proportions in **distr-3**.

Client no.	Proportion
1	10.6%
2	7.4%
3	12.0%
4	11.4%
5	8.8%
6	14.6%
7	10.0%
8	5.4%
9	10.2%
10	9.2%



954 Figure 4: Performance comparison for P_OC strategy with different data distributions.



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

972
973 **Importance Sampling.** Nguyen et al. (2020) introduced FOLB, a theoretically grounded client
974 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
975 follows:

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

$$981 IS_m^k = |\langle \nabla f_m(x^k, \xi_m^k), d^k \rangle|.$$

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

$$984 \pi^k = \left(\frac{IS_m^k}{\sum_{m'=1}^M IS_{m'}^k} \right)_{m=1}^M.$$

990 **Gradient-Norm-Based Sampling.** For the image classification problem on CIFAR-10 dataset, we
991 introduce an alternative client sampling strategy based on gradient norm sampling GNS Wang et al.
992 (2020b), which prioritizes clients whose local updates exhibit larger magnitudes. In particular:

993 1. At each communication round k , the server
994 estimates the relative importance of each client
995 m using the norm of its reported gradient
996 $\nabla f_m(w^k, \xi_m^k)$:

$$997 p_m^k = \frac{\|\nabla f_m(w^k, \xi_m^k)\|_2}{\sum_{m'=1}^M \|\nabla f_{m'}(w^k, \xi_{m'}^k)\|_2}.$$

999 2. Clients are then sampled with probabilities
1000 proportional to $\{p_m^k\}_{m=1}^M$, ensuring that those
1001 with larger gradient norms are selected more
1002 frequently:

$$1003 \pi^k = (p_m^k)_{m=1}^M.$$

1005 The obtained comparison results are presented
1006 in Figures 4, 6, 5, and 7.

1007 **ViT Fine-tuning.** To further assess the generalization and adaptability of our method, we conduct
1008 additional experiments involving the fine-tuning of a state-of-the-art Vision Transformer architecture
1009 FASTERViT (Hatamizadeh et al., 2023). The model, pre-trained on the large-scale IMAGENET21K
1010 dataset (Ridnik et al., 2021), comprises approximately 270M parameters and integrates hybrid
1011 hierarchical-attention mechanisms for efficient multi-scale feature learning. We fine-tune this model
1012 on the FOOD101 dataset (Bossard et al., 2014), a challenging benchmark consisting of 101,000
1013 images across 101 fine-grained food categories. This dataset presents significant visual complexity
1014 due to high class variation and subtle inter-class distinctions, making it particularly suitable for
1015 evaluating the scalability of our method.

1017 Table 4: Summary of training strategies used in additional experiments. Top and Rand denote the
1018 client selection rules, where the number indicates how many clients were selected for training.

1020 Epoch Strategy	1021 Round Strategy
1022 GNS (Top 3)	PoC (Top 1)
1023 FOLB (Top 3)	PoC (Top 1)
1024 PoC (Top 3)	Rand 1
1025 FOLB (Top 3)	Rand 1

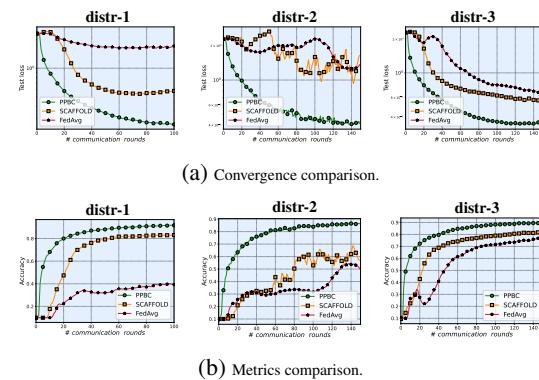


Figure 6: Performance comparison for FOLB strategy with different data distributions.

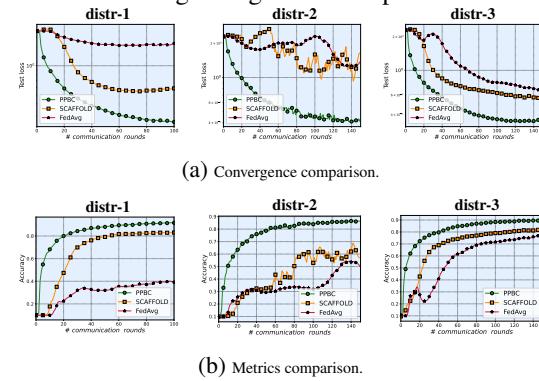


Figure 7: Performance comparison for GNS strategy with different data distributions.

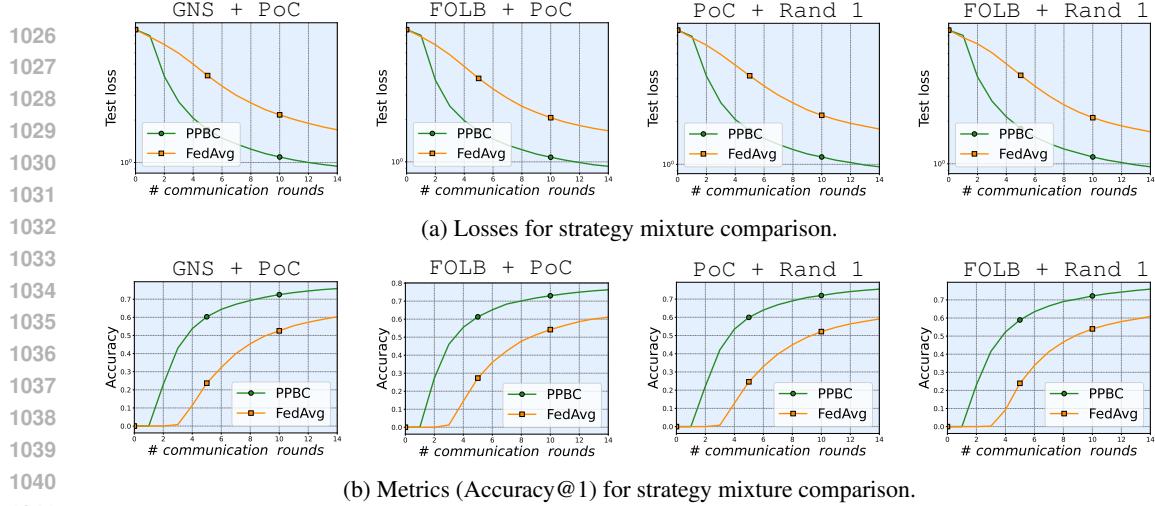


Figure 8: Performance comparison for combination of strategies on FASTERVIT fine-tuning.

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 $\tilde{\mathcal{R}}^{k,h}$ to change dynamically at each communication round k . The combinations of strategies are presented in Table 4. The performance validation results for each strategy mixture can be observed in Figure 8.

Ablation Study on Hyperparameters. Our framework admits a unifying interpretation: by setting $\theta = 0$ and disabling the client weighting mechanism, we recover the original baseline methods (FedAvg + any client sampling strategy). Consequently, by varying θ we can obtain various performance of Algorithm 1. Our method also utilizes another hyperparameter: the duration between global aggregations (length of the local epochs) H^k , modeled as a geometrically distributed random variable with parameter p . Our theoretical analysis imposes no constraints on p ; convergence guarantees hold for any choice, with rates explicitly dependent on this hyperparameter (see Theorems 3.5, 3.2, E.2, E.6). Next, we conduct an ablation study on both hyperparameters θ, p to quantify their impact on performance. Moreover, we demonstrate the empirical connection between θ and p , which correlates with our theoretical findings.

Firstly, we provide ablation study on θ . We fix $p = 0.2$ (yielding $H^k = 5$) and vary θ under the GNS client selection rule. Results are shown in Table 5.

Table 5: Ablation on θ with $H^k = 5$.

θ	Accuracy	Loss
0.05	0.88	0.35
0.10	0.90	0.31
0.15	0.93	0.21
0.20	0.89	0.32

We confirm our theoretical expectations: excessively small values of θ do not allow for effectively accounting for the clients' history ($\theta = 0$ corresponds to FedAvg), while large values disproportionately increases the contribution of gradient surrogates that become outdated after an epoch. However, there exists a wide interval within which the method do not lose much quality compared to optimal θ value.

1080 Next, we fix $\theta = 0.2$ and vary p (i.e., the expected epoch size $H^k = 1/p$), with results in Table 6.
1081

1082 Table 6: Ablation on local epoch size with $\theta = 0.2$.
1083

H^k	Accuracy	Loss
1	0.81	0.38
3	0.91	0.23
5	0.89	0.32
7	0.82	0.39

1091 For $\theta = 0.15$, the optimal local epoch size is $H^k = 5$ (see Table 6), while for $\theta = 0.2$, the optimal
1092 value decreases to $H^k = 3$. This finding is in complete agreement with theoretical expectations:
1093 bigger values of θ require fewer number of local steps to achieve optimal convergence.
1094

1095 **Ablation Study on Convergence.** In this para-
1096 graph, we emphasize that the proposed Algo-
1097 rithm 1 maintains similar convergence behavior
1098 across all combinations of the considered stra-
1099 tegies (see Figure 9). This result is obtained by
1100 gradient compensation technique incorporated
1101 in our method. Thus, a biases that appear due to
1102 applying client sampling strategies are equally
1103 mitigated by our algorithm.
1104

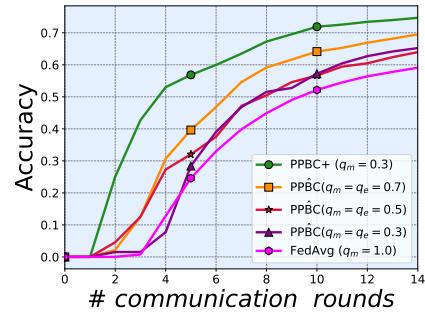
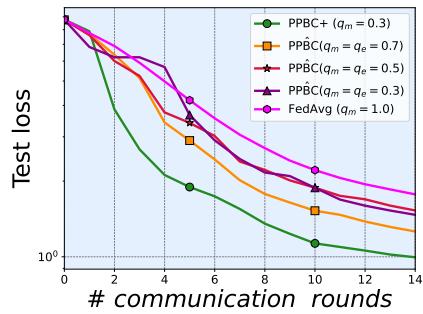
1105 **PPBC does not require a fixed aggregation**

1106 **round.** Our algorithms PPBC and PPBC+ have one limitation: they require transmitting all
1107 accumulated surrogates once per epoch. For this reason, we conduct an experimental study ($\hat{\text{PPBC}}$) in
1108 which we remove the requirement that all devices must send their information every fixed number of
1109 iterations.

1110 We introduced an additional mechanism: at the moment of full aggregation, a client may choose not
1111 to send the surrogate it accumulated during the epoch. This is modeled similarly to PPBC+, using
1112 a Bernoulli random variable with a new hyperparameter q_e . In other words, any client may fail to
1113 provide its surrogate during the full aggregation step. Consequently, line 20 of the Algorithm 2 is
1114 modified to the following block:

$$1115 \text{generate } \eta_m^k \sim \mathcal{Q}_e, \quad \text{if } \eta_m^k = 1 : \text{send } g_m^{k, H_k} \text{ to the server}$$

1116 We conducted experiments (see Figure 10) for different values of both q_m and q_e , and compared our
1117 results with standard PPBC+ (Algorithm 2) using $q_m = 0.3$, as well as with the baseline FedAvg
1118 under pathological data heterogeneity. As expected, the new algorithm performs worse than PPBC+
1119 with full aggregation, yet it still consistently outperforms FedAvg.

1132 Figure 10: Test loss and test accuracy of $\hat{\text{PPBC}}$, PPBC+, and FedAvg on FASTERViT fine-tuning on
1133 FOOD101.

1134 **C GENERAL STATEMENTS**
 1135

1136 **Notation.** In the work we use the following notation. $x^{k,h} \in \mathbb{R}^d$ is the vector of model's parameters
 1137 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
 1138 $x \in \mathbb{R}^d$, $\nabla f_m(x, \xi) \in \mathbb{R}^d$ denotes the stochastic gradient at the point $x \in \mathbb{R}^d$ with respect to
 1139 stochastic realization ξ .

1140 For a random vector $x \in \mathbb{R}^d$ and stochasticity ξ we denote $\mathbb{E}[x]$ is the expected value of x and $\mathbb{E}_\xi[x]$
 1141 as the conditioned expected value with the respect to ξ .

1142 We use $\|x\| = \sqrt{\sum_{i=1}^d x_i^2}$ as l_2 -norm of the vector $x \in \mathbb{R}^d$ and $\langle x, y \rangle = \sum_{i=1}^d$ represents the scalar
 1143 product of vectors $x, y \in \mathbb{R}^d$.
 1144

1145 We use *number of devices communications* (device to server communications) as the metric. This
 1146 choice arises from the recognition that the number of rounds of communication is insufficient to
 1147 adequately compare distributed methods. For example, this limitation becomes evident when the
 1148 nodes operate asynchronously. In this case, the more appropriate metric is the total number of
 1149 communications rather than the number of rounds.
 1150

1151 **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
 1152 2.1, 2.2(b), $\varphi(\cdot)$ is under Assumption 2.2(b). Then,

$$1153 \|\nabla f(x) - \nabla f(y)\|^2 \leq 2L(f(x) - f(y) - \langle \nabla f(y), x - y \rangle), \quad (\text{Lip})$$

$$1154 \langle x, y \rangle \leq \frac{\beta}{2} \|x\|^2 + \frac{1}{2\beta} \|y\|^2, \quad (\text{Fen})$$

$$1155 \left\| \sum_{i=1}^n a_i \right\|^2 \leq n \sum_{i=1}^n \|a_i\|^2, \quad (\text{CS})$$

$$1156 \varphi \left(\frac{\sum_{i=1}^n w_i a_i}{\sum_{i=1}^n a_i} \right) \leq \frac{\sum_{i=1}^n a_i \varphi(x_i)}{\sum_{i=1}^n a_i}. \quad (\text{Jen})$$

1157 **Lemma C.1** ((Allen-Zhu, 2018)). Given sequence $D_0, D_1, \dots, D_N \in \mathbb{R}$, where $N \in \text{Geom}(p)$.
 1158 Then,

$$1159 \mathbb{E}_N[D_{N-1}] = pD_0 + (1-p)\mathbb{E}_N[D_N].$$

1160 **D PROOFS FOR ALGORITHM 1**

1161 **Lemma D.1 (Lemma 3.1).** Suppose Assumptions 2.3, 2.4 hold. Then for Algorithm 1 it implies that

$$1162 \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|g^{k,H^k}\|^2 \leq \frac{24(1-\theta)^2 \alpha(\delta_1 + 1)}{p^2} \mathbb{E}_{H^k} \|\nabla f(x^{k,H^k})\|^2 + \frac{48(1-\theta)^2 \alpha \delta_2}{p^2} \\ 1163 + \frac{24(1-\theta)^2 \alpha \sigma^2}{Mp^2}.$$

1164 *Proof.* Let us start with the following estimate:

$$1165 \|g^{k,h+1}\|^2 = \left\| g^{k,h} + (1-\theta) \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\ 1166 \stackrel{(\text{Fen})}{\leq} (1+c) \|g^{k,h}\|^2 \\ 1167 + \left(1 + \frac{1}{c} \right) (1-\theta)^2 \left\| \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2, \quad (4)$$

1168 where c is defined below. Let us estimate the last term and obtain

$$1169 \left\| \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2$$

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

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

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

We apply equation CS to the first term and identically transform the second and third terms:

$$\begin{aligned}
& \left\| \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\
& \leq 4 \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right)^2 \sum_{m=1}^M \|\nabla f_m(x^{k,h}) - \nabla f(x^{k,h})\|^2 \\
& \quad + 4 \left(\sum_{m=1}^M (\tilde{\pi}_m^{k,h} - \pi_m^k) \right)^2 \|\nabla f(x^{k,h})\|^2 \\
& \quad + 2 \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right)^2 \|\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})\|^2 \\
& \quad + 4 \sum_{i \neq j} \left(\frac{1}{M} - \tilde{\pi}_i^{k,h} \right) \left(\frac{1}{M} - \tilde{\pi}_j^{k,h} \right) \\
& \quad \cdot \langle \nabla f_i(x^{k,h}, \xi_i^{k,h}) - \nabla f_i(x^{k,h}), \nabla f_j(x^{k,h}, \xi_j^{k,h}) - \nabla f_j(x^{k,h}) \rangle \\
& \stackrel{\text{As. 2.3}}{\leq} 4M \left(\delta_1 \|\nabla f(x^{k,h})\|^2 + \delta_2 \right) \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right)^2 + 4 \|\nabla f(x^{k,h})\|^2
\end{aligned}$$

$$\begin{aligned}
& +2 \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right)^2 \|\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})\|^2 \\
& +4 \sum_{i \neq j} \left(\frac{1}{M} - \tilde{\pi}_i^{k,h} \right) \left(\frac{1}{M} - \tilde{\pi}_j^{k,h} \right) \cdot \\
& \cdot \left\langle \nabla f_i(x^{k,h}, \xi_i^{k,h}) - \nabla f_i(x^{k,h}), \nabla f_j(x^{k,h}, \xi_j^{k,h}) - \nabla f_j(x^{k,h}) \right\rangle,
\end{aligned}$$

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 $\sum_{m=1}^M \left(\tilde{\pi}_m^{k,h} - \pi_m^k \right) \leq 1$.

Taking expectation on $\xi_m^{k,h}$ and using Assumption 2.4, we have

$$\begin{aligned}
\mathbb{E}_{\xi_m^{k,h}} \left\| \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 & \leq 4M\delta_1 \|\nabla f(x^{k,h})\|^2 \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right)^2 \\
& + 4 \|\nabla f(x^{k,h})\|^2 \\
& + 4M\delta_2 \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right)^2 \\
& + 2\sigma^2 \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h} \right)^2, \tag{5}
\end{aligned}$$

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

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

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

We can transform equation 5 into

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

Substituting equation 7 into equation 4, we have

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

Enrolling a recursion, we get

$$\mathbb{E}_{\xi_m^{k,0}} \dots \mathbb{E}_{\xi_m^{k,h}} \|g^{k,h+1}\|^2 \leq 4 \left(1 + \frac{1}{c} \right) (1-\theta)^2 \alpha(\delta_1 + 1) \sum_{i=0}^h (1+c)^{h-i} \|\nabla f(x^{k,i})\|^2$$

$$\begin{aligned}
& +4 \left(1 + \frac{1}{c}\right) (1 - \theta)^2 \alpha \delta_2 \sum_{i=0}^h (1 + c)^{h-i} \\
& +2 \left(1 + \frac{1}{c}\right) (1 - \theta)^2 \frac{\alpha}{M} \sigma^2 \sum_{i=0}^h (1 + c)^{h-i}.
\end{aligned} \tag{8}$$

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

$$\begin{aligned}
\mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \|g^{k,H^k}\|^2 &= \sum_{j \geq 0} p(1-p)^j \mathbb{E}_{\xi_m^{k,0}} \dots \mathbb{E}_{\xi_m^{k,j-1}} \|g^{k,j}\|^2 \\
&\stackrel{(8)}{\leq} 4 \left(1 + \frac{1}{c}\right) (1 - \theta)^2 \alpha (\delta_1 + 1) \cdot \\
&\quad \cdot \sum_{j \geq 0} p(1-p)^j \sum_{i=0}^{j-1} (1 + c)^{j-i-1} \|\nabla f(x^{k,i})\|^2 \\
&\quad +2 \left(1 + \frac{1}{c}\right) (1 - \theta)^2 \frac{\alpha}{M} (\sigma^2 + 2M\delta_2) \cdot \\
&\quad \cdot \sum_{j \geq 0} p(1-p)^j \sum_{i=0}^{j-1} (1 + c)^{j-i-1}.
\end{aligned} \tag{9}$$

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

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

Additionally, we have

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

Combining this estimates with equation 9 we obtain the result of the lemma:

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

$$+ \frac{24(1-\theta)^2\alpha\sigma^2}{Mp^2}.$$

□

D.1 PROOF FOR NON-CONVEX CASE

Theorem D.2 (Theorem 3.2). *Suppose Assumptions 2.1, 2.2(a), 2.3, 2.4 hold. Then for Algorithm 1 with $\theta \leq \frac{\gamma L p^2}{2}$ and $\gamma \leq \frac{p}{384L\alpha(\delta_1+1)}$ it implies that*

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

Proof. We start with the definition of virtual sequence:

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

It is followed by

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

Assumption 2.1 implies

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

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

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

Note that

$$\tilde{x}^{k,h} \stackrel{(12)}{=} x^{k,h} - \gamma g^{k,h}$$

$$1404 \quad \text{Line 12} \quad x^{k,h} - \gamma \left(g^{k,h-1} + (1-\theta) \sum_{m=1}^M \left(\frac{1}{M} - \tilde{\pi}_m^{k,h-1} \right) \nabla f(x^{k,h-1}, \xi_m^{k,h-1}) \right).$$

1407 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,
1408 equation 14 transforms into
1409

$$1410 \quad \mathbb{E}_{\xi_m^{k,h}} [f(\tilde{x}^{k,h+1})] \leq f(\tilde{x}^{k,h}) - \gamma \theta \langle \nabla f(\tilde{x}^{k,h}), g^{k-1,H^{k-1}} \rangle \\ 1411 \quad - \gamma(1-\theta) \langle \nabla f(\tilde{x}^{k,h}), \nabla f(x^{k,h}) \rangle \\ 1412 \quad + \frac{\gamma^2 L(1-\theta)}{2} \mathbb{E}_{\xi_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\ 1413 \quad + \frac{\gamma^2 L\theta}{2} \|g^{k-1,H^{k-1}}\|^2 \\ 1414 \quad \stackrel{(CS)}{\leq} f(\tilde{x}^{k,h}) - \gamma \theta \langle \nabla f(\tilde{x}^{k,h}), g^{k-1,H^{k-1}} \rangle \\ 1415 \quad - \gamma(1-\theta) \langle \nabla f(\tilde{x}^{k,h}), \nabla f(x^{k,h}) \rangle \\ 1416 \quad + \gamma^2 L(1-\theta) \mathbb{E}_{\xi_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M (\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})) \right\|^2 \\ 1417 \quad + \gamma^2 L(1-\theta) \|\nabla f(x^{k,h})\|^2 \\ 1418 \quad + \frac{\gamma^2 L\theta}{2} \|g^{k-1,H^{k-1}}\|^2. \quad (15)$$

1420 Now we pay attention to the following term:
1421

$$1422 \quad \mathbb{E}_{\xi_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M (\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})) \right\|^2 \\ 1423 \quad \stackrel{(i)}{=} \frac{1}{M^2} \sum_{m=1}^M \mathbb{E}_{\xi_m^{k,h}} \|\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})\|^2 \\ 1424 \quad + \frac{2}{M^2} \sum_{i \neq j} \left\langle \mathbb{E}_{\xi_i^{k,h}} [\nabla f_i(x^{k,h}, \xi_i^{k,h}) - \nabla f_i(x^{k,h})], \mathbb{E}_{\xi_j^{k,h}} [\nabla f_j(x^{k,h}, \xi_j^{k,h}) - \nabla f_j(x^{k,h})] \right\rangle \\ 1425 \quad \stackrel{\text{As. 2.4}}{\leq} \frac{1}{M} \sigma^2,$$

1426 where (i) is correct, since $\xi_i^{k,h}$ and $\xi_j^{k,h}$ are independent. Substituting this estimate into equation 15,
1427 we have
1428

$$1429 \quad \mathbb{E}_{\xi_m^{k,h}} [f(\tilde{x}^{k,h+1})] \leq f(\tilde{x}^{k,h}) - \gamma \theta \langle \nabla f(\tilde{x}^{k,h}), g^{k-1,H^{k-1}} \rangle \\ 1430 \quad - \gamma(1-\theta) \langle \nabla f(\tilde{x}^{k,h}), \nabla f(x^{k,h}) \rangle \\ 1431 \quad + \gamma^2 L(1-\theta) \|\nabla f(x^{k,h})\|^2 + \frac{\gamma^2 L\theta}{2} \|g^{k-1,H^{k-1}}\|^2 \\ 1432 \quad + \frac{\gamma^2 L(1-\theta)\sigma^2}{M}. \quad (16)$$

1433 Let us estimate the scalar products separately.
1434

$$1435 \quad - \gamma(1-\theta) \langle \nabla f(\tilde{x}^{k,h}), \nabla f(x^{k,h}) \rangle = -\frac{\gamma(1-\theta)}{2} \|\nabla f(\tilde{x}^{k,h})\|^2 - \frac{\gamma(1-\theta)}{2} \|\nabla f(x^{k,h})\|^2 \\ 1436 \quad + \frac{\gamma(1-\theta)}{2} \|\nabla f(\tilde{x}^{k,h}) - \nabla f(x^{k,h})\|^2$$

$$\begin{aligned}
& \stackrel{\text{As. 2.1}}{\leq} -\frac{\gamma(1-\theta)}{2} \|\nabla f(\tilde{x}^{k,h})\|^2 - \frac{\gamma(1-\theta)}{2} \|\nabla f(x^{k,h})\|^2 \\
& \quad + \frac{\gamma L^2(1-\theta)}{2} \|\tilde{x}^{k,h} - x^{k,h}\|^2 \\
& \stackrel{(12)}{=} -\frac{\gamma(1-\theta)}{2} \|\nabla f(\tilde{x}^{k,h})\|^2 - \frac{\gamma(1-\theta)}{2} \|\nabla f(x^{k,h})\|^2 \\
& \quad + \frac{\gamma^3 L^2(1-\theta)}{2} \|g^{k,h}\|^2, \\
& -\gamma\theta \left\langle \nabla f(\tilde{x}^{k,h}), g^{k-1,H^{k-1}} \right\rangle \stackrel{(Fen)}{\leq} \frac{\gamma\theta}{2} \|\nabla f(\tilde{x}^{k,h})\|^2 + \frac{\gamma\theta}{2} \|g^{k-1,H^{k-1}}\|^2.
\end{aligned}$$

Combining it with equation 16, we have

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

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

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

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

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

$$\begin{aligned}
& + \frac{\gamma\theta(\gamma L + 1)}{2} \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0} \dots \xi_m^{k-1,H^{k-1}-1}} \|g^{k-1,H^{k-1}}\|^2 \\
& + \frac{\gamma^2 L(1-\theta)\sigma^2}{M}.
\end{aligned}$$

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 stochastic values.

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

We use Lemma 3.1 to estimate $\|g^{k,H^k}\|^2$ and $\|g^{k-1,H^{k-1}}\|^2$. We obtain

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

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

$$\begin{aligned}
& \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0} \dots \xi_m^{k-1,H^{k-1}-1}} \|g^{k-1,H^{k-1}}\|^2 \\
& \leq \frac{24(1-\theta)^2\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \|\nabla f(x^{k-1,H^{k-1}})\|^2 + \frac{48(1-\theta)^2\alpha\delta_2}{p^2} + \frac{24(1-\theta)^2\alpha\sigma^2}{Mp^2} \\
& \stackrel{(CS)}{\leq} \frac{48(1-\theta)\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \|\nabla f(x^{k-1,H^{k-1}}) - \nabla f(\tilde{x}^{k-1,H^{k-1}})\|^2 \\
& + \frac{48(1-\theta)^2\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \|\nabla f(\tilde{x}^{k-1,H^{k-1}})\|^2 + \frac{48(1-\theta)^2\alpha\delta_2}{p^2} + \frac{24(1-\theta)^2\alpha\sigma^2}{Mp^2} \\
& \stackrel{\text{As. 2.1}}{\leq} \frac{48L^2(1-\theta)^2\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \|x^{k-1,H^{k-1}} - \tilde{x}^{k-1,H^{k-1}}\|^2 \\
& + \frac{48(1-\theta)^2\alpha(\delta_1+1)}{p^2} \|\nabla f(x^{k,0})\|^2 + \frac{48(1-\theta)^2\alpha\delta_2}{p^2} + \frac{24(1-\theta)^2\alpha\sigma^2}{Mp^2} \\
& \stackrel{(12)}{=} \frac{48\gamma^2 L^2(1-\theta)^2\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \|g^{k-1,H^{k-1}}\|^2 \\
& + \frac{48(1-\theta)^2\alpha(\delta_1+1)}{p^2} \|\nabla f(x^{k,0})\|^2 + \frac{48(1-\theta)^2\alpha\delta_2}{p^2} + \frac{24(1-\theta)^2\alpha\sigma^2}{Mp^2}.
\end{aligned}$$

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

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

Now we substitute equation 18 and equation 19 into equation 17:

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

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

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

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 choice of θ and γ , consequently this choice is correct. In that way, we obtain

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

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,

$$\frac{\gamma(1-\theta)}{8} \mathbb{E} \left\| \nabla f(x^{k,0}) \right\|^2 \leq \mathbb{E} \left[f(x^{k,0}) - f(x^{k+1,0}) \right] + \frac{48\gamma L\alpha\delta_2}{p} + \frac{24\gamma^2L^2\alpha\delta_2}{p^3}$$

$$1620 \quad + \frac{25\gamma^2 L\alpha\sigma^2}{Mp} + \frac{12\gamma^3 L^2\alpha\sigma^2}{Mp^3}.$$

1622 Averaging over all epochs, we obtain the result of the theorem:
1623

$$1624 \quad \begin{aligned} 1625 \quad \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(x^{k,0})\|^2 &\leq \frac{8(f(x^{0,0}) - \mathbb{E}[f(x^{K,0})])}{\gamma(1-\theta)K} + \frac{384\gamma L\alpha\delta_2}{p(1-\theta)} + \frac{192\gamma^2 L^2\alpha\delta_2}{p^3(1-\theta)} \\ 1626 \quad &+ \frac{200\gamma L\alpha\sigma^2}{Mp(1-\theta)} + \frac{96\gamma^2 L^2\alpha\sigma^2}{Mp^3(1-\theta)} \\ 1627 \quad &\leq \frac{16(f(x^{0,0}) - f(x^*))}{\gamma K} + \frac{768\gamma L\alpha\delta_2}{p} + \frac{384\gamma^2 L^2\alpha\delta_2}{p^3} \\ 1628 \quad &+ \frac{400\gamma L\alpha\sigma^2}{Mp} + \frac{192\gamma^2 L^2\alpha\sigma^2}{Mp^3}. \end{aligned}$$

1629 \square

1630 **Corollary D.3 (Corollary 3.3).** *Under conditions of Theorem 3.2 Algorithm 1 with fixed rules
1631 $\hat{\mathcal{R}}^k \equiv \hat{\mathcal{R}}^{k,h} \equiv \mathcal{R}$ needs*

$$1632 \quad \mathcal{O}\left(\frac{M}{C}\left(\frac{\Delta L\alpha\delta_1}{\varepsilon^2} + \frac{\Delta L\alpha\delta_2}{\varepsilon^4} + \frac{\Delta L\alpha\sigma^2}{M\varepsilon^4}\right)\right) \text{ epochs and}$$

$$1633 \quad \mathcal{O}\left(M\frac{M}{C}\left(\frac{\Delta L\alpha\delta_1}{\varepsilon^2} + \frac{\Delta L\alpha\delta_2}{\varepsilon^4} + \frac{\Delta L\alpha\sigma^2}{M\varepsilon^4}\right)\right) \text{ number of devices communications}$$

1634 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
1635 of devices participating in each epoch.

1636 *Proof.* Using the result of Theorem 3.2, we choose

$$1637 \quad \gamma \leq \min \left\{ \frac{p}{384L\alpha(\delta_1 + 1)}, \frac{\sqrt{(f(x^{0,0}) - f(x^*))p}}{4\sqrt{3L\alpha\delta_2 K}}, \frac{\sqrt[3]{(f(x^{0,0}) - f(x^*))p}}{2\sqrt[3]{3L^2\alpha\delta_2 K}}, \right. \\ 1638 \quad \left. \frac{\sqrt{(f(x^{0,0}) - f(x^*))Mp}}{5\sqrt{L\alpha\sigma^2 K}}, \frac{\sqrt[3]{(f(x^{0,0}) - f(x^*))Mp}}{\sqrt[3]{12L^2\alpha\sigma^2 K}} \right\}.$$

1639 Thus, we need

$$1640 \quad \mathcal{O}\left(\frac{(f(x^{0,0}) - f(x^*))L\alpha\delta_1}{p\varepsilon^2} + \frac{(f(x^{0,0}) - f(x^*))L\alpha\delta_2}{p\varepsilon^4} + \frac{(f(x^{0,0}) - f(x^*))L\alpha\sigma^2}{Mp\varepsilon^4} \right. \\ 1641 \quad \left. + \frac{(f(x^{0,0}) - f(x^*))L\sqrt{\alpha\delta_2}}{p^{\frac{3}{2}}\varepsilon^3} + \frac{(f(x^{0,0}) - f(x^*))L\sqrt{\alpha\sigma}}{\sqrt{Mp^{\frac{3}{2}}\varepsilon^3}} \right)$$

1642 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
1643 terms in the estimate in a magnitude smaller, than the second and third accordingly,
1644 we can ignore them. The length of the epoch $H \in \text{Geom}(p)$, Algorithm 1 requires
1645 $\mathcal{O}\left(\frac{(f(x^{0,0}) - f(x^*))L\alpha\delta_1}{p^2\varepsilon^2} + \frac{(f(x^{0,0}) - f(x^*))L\alpha\delta_2}{p^2\varepsilon^4} + \frac{(f(x^{0,0}) - f(x^*))L\alpha\sigma^2}{Mp^2\varepsilon^4}\right)$ communication rounds.

1646 Next we mention that at each communication round we communicate with C devices, thus, number of
1647 communications is $\mathcal{O}\left(C\frac{(f(x^{0,0}) - f(x^*))L\alpha\delta_1}{p^2\varepsilon^2} + C\frac{(f(x^{0,0}) - f(x^*))L\alpha\delta_2}{p^2\varepsilon^4} + C\frac{(f(x^{0,0}) - f(x^*))L\alpha\sigma^2}{Mp^2\varepsilon^4}\right)$.

1648 Taking $p = \frac{C}{M}$, we have the result of the corollary. The choice of p is motivated by the fact
1649 that we perform $\frac{1}{p}C + M$ communications per-epoch, and established p is the minimal, which
1650 delivers $\mathcal{O}(M)$ communications at each epoch. This is also the reason for the additional factor M in
1651 the estimate on communications. \square

1674 **Corollary D.4.** *Under conditions of Theorem 3.2 Algorithm 1 needs*

$$1675 \quad \mathcal{O} \left(\frac{M}{\min_{k,h} C^{k,h}} \left(\frac{\Delta L \alpha \delta_1}{\varepsilon^2} + \frac{\Delta L \alpha \delta_2}{\varepsilon^4} + \frac{\Delta L \alpha \sigma^2}{M \varepsilon^4} \right) \right) \text{ epochs and}$$

$$1676 \quad \mathcal{O} \left(M \left(\frac{M}{\min_{k,h} C^{k,h}} \right)^2 \left(\frac{\Delta L \alpha \delta_1}{\varepsilon^2} + \frac{\Delta L \alpha \delta_2}{\varepsilon^4} + \frac{\Delta L \alpha \sigma^2}{M \varepsilon^4} \right) \right) \text{ number of devices communications}$$

1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2599
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2639
2640
2641
2642
2643
2644
2645
2646
2647
2648
2649
2649
2650
2651
2652
2653
2654
2655
2656
2657
2658
2659
2659
2660
2661
2662
2663
2664
2665
2666
2667
2668
2669
2669
2670
2671
2672
2673
2674
2675
2676
2677
2678
2679
2679
2680
2681
2682
2683
2684
2685
2686
2687
2688
2689
2690
2691
2692
2693
2694
2695
2696
2697
2698
2699
2699
2700
2701
2702
2703
2704
2705
2706
2707
2708
2709
2710
2711
2712
2713
2714
2715
2716
2717
2718
2719
2720
2721
2722
2723
2724
2725
2726
2727
2728
2729
2729
2730
2731
2732
2733
2734
2735
2736
2737
2738
2739
2739
2740
2741
2742
2743
2744
2745
2746
2747
2748
2749
2749
2750
2751
2752
2753
2754
2755
2756
2757
2758
2759
2759
2760
2761
2762
2763
2764
2765
2766
2767
2768
2769
2769
2770
2771
2772
2773
2774
2775
2776
2777
2778
2779
2779
2780
2781
2782
2783
2784
2785
2786
2787
2788
2789
2789
2790
2791
2792
2793
2794
2795
2796
2797
2798
2799
2799
2800
2801
2802
2803
2804
2805
2806
2807
2808
2809
2809
2810
2811
2812
2813
2814
2815
2816
2817
2818
2819
2819
2820
2821
2822
2823
2824
2825
2826
2827
2828
2829
2829
2830
2831
2832
2833
2834
2835
2836
2837
2838
2839
2839
2840
2841
2842
2843
2844
2845
2846
2847
2848
2849
2849
2850
2851
2852
2853
2854
2855
2856
2857
2858
2859
2859
2860
2861
2862
2863
2864
2865
2866
2867
2868
2869
2869
2870
2871
2872
2873
2874
2875
2876
2877
2878
2879
2879
2880
2881
2882
2883
2884
2885
2886
2887
2888
2889
2889
2890
2891
2892
2893
2894
2895
2896
2897
2898
2899
2899
2900
2901
2902
2903
2904
2905
2906
2907
2908
2909
2909
2910
2911
2912
2913
2914
2915
2916
2917
2918
2919
2919
2920
2921
2922
2923
2924
2925
2926
2927
2928
2929
2929
2930
2931
2932
2933
2934
2935
2936
2937
2938
2939
2939
2940
2941
2942
2943
2944
2945
2946
2947
2948
2949
2949
2950
2951
2952
2953
2954
2955
2956
2957
2958
2959
2959
2960
2961
2962
2963
2964
2965
2966
2967
2968
2969
2969
2970
2971
2972
2973
2974
2975
2976
2977
2978
2979
2979
2980
2981
2982
2983
2984
2985
2986
2987
2988
2989
2989
2990
2991
2992
2993
2994
2995
2996
2997
2998
2999
2999
3000
3001
3002
3003
3004
3005
3006
3007
3008
3009
3009
3010
3011
3012
3013
3014
3015
3016
3017
3018
3019
3019
3020
3021
3022
3023
3024
3025
3026
3027
3028
3029
3029
3030
3031
3032
3033
3034
3035
3036
3037
3038
3039
3039
3040
3041
3042
3043
3044
3045
3046
3047
3048
3049
3049
3050
3051
3052
3053
3054
3055
3056
3057
3058
3059
3059
3060
3061
3062
3063
3064
3065
3066
3067
3068
3069
3069
3070
3071
3072
3073
3074
3075
3076
3077
3078
3079
3079
3080
3081
3082
3083
3084
3085
3086
3087
3088
3089
3089
3090
3091
3092
3093
3094
3095
3096
3097
3098
3099
3099
3100
3101
3102
3103
3104
3105
3106
3107
3108
3109
3109
3110
3111
3112
3113
3114
3115
3116
3117
3118
3119
3119
3120
3121
3122
3123
3124
3125
3126
3127
3128
3129
3129
3130
3131
3132
3133
3134
3135
3136
3137
3138
3139
3139
3140
3141
3142
3143
3144
3145
3146
3147
3148
3149
3149
3150
3151
3152
3153
3154
3155
3156
3157
3158
3159
3159
3160
3161
3162
3163
3164
3165
3166
3167
3168
3169
3169
3170
3171
3172
3173
3174
3175
3176
3177
3178
3179
3179
3180
3181
3182
3183
3184
3185
3186
3187
3188
3189
3189
3190
3191
3192
3193
3194
3195
3196
3197
3198
3199
3199
3200
3201
3202
3203
3204
3205
3206
3207
3208
3209
3209
3210
3211
3212
3213
3214
3215
3216
3217
3218
3219
3219
3220
3221
3222
3223
3224
3225
3226
3227
3228
3229
3229
3230
3231
3232
3233
3234
3235
3236
3237
3238
3239
3239
3240
3241
3242
3243
3244
3245
3246
3247
3248
3249
3249
3250
3251
3252
3253
3254
3255
3256
3257
3258
3259
3259
3260
3261
3262
3263
3264
3265
3266
3267
3268
3269
3269
3270
3271
3272
3273
3274
3275
3276
3277
3278
3279
3279
3280
3281
3282
3283
3284
3285
3286
3287
3288
3289
3289
3290
3291
3292
3293
3294
3295
3296
3297
3298
3299
3299
3300
3301
3302
3303
3304
3305
3306
3307
3308
3309
3309
3310
3311
3312
3313
3314
3315
3316
3317
3318
3319
3319
3320
3321
3322
3323
3324
3325
3326
3327
3328
3329
3329
3330
3331
3332
3333
3334
3335
3336
3337
3338
3339
3339
3340
3341
3342
3343
3344
3345
3346
3347
3348
3349
3349
3350
3351
3352
3353
3354
3355
3356
3357
3358
3359
3359
3360
3361
3362
3363
3364
3365
3366
3367
3368
3369
3369
3370
3371
3372
3373
3374
3375
3376
3377
3378
3379
3379
3380
3381
3382
3383
3384
3385
3386
3387
3388
3389
3389
3390
3391
3392
3393
3394
3395
3396
3397
3398
3399
3399
3400
3401
3402
340

1728 D.2 PROOF FOR STRONGLY-CONVEX CASE
 1729

1730 **Theorem D.6 (Theorem 3.5).** *Suppose Assumptions 2.1, 2.2(b), 2.3, 2.4 hold. Then for Algorithm 1
 1731 with $\theta \leq \frac{p\gamma\mu}{4}$ and $\gamma \leq \frac{p^2}{96L\alpha(\delta_1+1)}$ it implies that*

$$1732 \mathbb{E} \|x^{K,0} - x^*\|^2 \leq \left(1 - \frac{\gamma\mu}{8}\right)^K \|x^{0,0} - x^*\|^2 + \frac{8\gamma\alpha}{\mu p^3} \left(144\delta_2 + \frac{74\sigma^2}{M}\right).$$

1735 *Proof.* We start with the definition of virtual sequence:

$$1737 \widetilde{x}^{k,h} = x^{k,h} - \gamma \sum_{m=1}^M g_m^{k,h}. \quad (20)$$

1739 It is followed by

$$\begin{aligned} 1741 \widetilde{x}^{k,h+1} &= x^{k,h+1} - \gamma \sum_{m=1}^M g_m^{k,h+1} \\ 1742 &= x^{k,h} - \gamma \left[(1-\theta) \sum_{m=1}^M \widetilde{\pi}_m^{k,h} \nabla f_m(x^{k,h}, \xi_m^{k,h}) + \theta g^{k-1, H^{k-1}} \right] \\ 1743 &\quad - \gamma \sum_{m=1}^M g_m^{k,h} - \gamma (1-\theta) \sum_{m=1}^M \left(\frac{1}{M} - \widetilde{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \\ 1744 &= \widetilde{x}^{k,h} - \gamma \left[(1-\theta) \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) + \theta g^{k-1, H^{k-1}} \right]. \end{aligned} \quad (21)$$

1752 We use this to write a descent:

$$\begin{aligned} 1753 \|\widetilde{x}^{k,h+1} - x^*\|^2 &= \|\widetilde{x}^{k,h} - x^*\|^2 + 2 \langle \widetilde{x}^{k,h} - x^*, \widetilde{x}^{k,h+1} - \widetilde{x}^{k,h} \rangle + \|\widetilde{x}^{k,h+1} - \widetilde{x}^{k,h}\|^2 \\ 1754 &\stackrel{(21)}{=} \|\widetilde{x}^{k,h} - x^*\|^2 - 2\gamma\theta \langle \widetilde{x}^{k,h} - x^*, g^{k-1, H^{k-1}} \rangle \\ 1755 &\quad - 2\gamma(1-\theta) \left\langle \widetilde{x}^{k,h} - x^*, \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ 1756 &\quad + \gamma^2 \left\| \theta g^{k-1, H^{k-1}} + (1-\theta) \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\ 1757 &\stackrel{(Jen)}{\leq} \|\widetilde{x}^{k,h} - x^*\|^2 - 2\gamma\theta \langle \widetilde{x}^{k,h} - x^*, g^{k-1, H^{k-1}} \rangle \\ 1758 &\quad - 2\gamma(1-\theta) \left\langle \widetilde{x}^{k,h} - x^k, \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ 1759 &\quad - 2\gamma(1-\theta) \left\langle x^{k,h} - x^*, \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ 1760 &\quad + \gamma^2 \theta \left\| g^{k-1, H^{k-1}} \right\|^2 + \gamma^2 (1-\theta) \left\| \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2. \end{aligned}$$

1761 Taking the expectation over $\xi_m^{k,h}$, we have

$$\begin{aligned} 1762 \mathbb{E}_{\xi_m^{k,h}} \|\widetilde{x}^{k,h+1} - x^*\|^2 &\leq \mathbb{E}_{\xi_m^{k,h}} \|\widetilde{x}^{k,h} - x^*\|^2 - 2\gamma\theta \mathbb{E}_{\xi_m^{k,h}} \langle \widetilde{x}^{k,h} - x^*, g^{k-1, H^{k-1}} \rangle \\ 1763 &\quad - 2\gamma(1-\theta) \mathbb{E}_{\xi_m^{k,h}} \left\langle \widetilde{x}^{k,h} - x^k, \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ 1764 &\quad - 2\gamma(1-\theta) \mathbb{E}_{\xi_m^{k,h}} \left\langle x^{k,h} - x^*, \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ 1765 &\quad + \gamma^2 \theta \left\| g^{k-1, H^{k-1}} \right\|^2 + \gamma^2 (1-\theta) \left\| \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2. \end{aligned}$$

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

1787 Mention that

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

1793 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,
1794 equation 22 transforms into

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

1815 Now we pay attention to the following term:
1816

$$\begin{aligned}
& \mathbb{E}_{\xi_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M (\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})) \right\|^2 \\
& \stackrel{(i)}{=} \frac{1}{M^2} \sum_{m=1}^M \mathbb{E}_{\xi_m^{k,h}} \|\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})\|^2 \\
& \quad + \frac{2}{M^2} \sum_{i \neq j} \left\langle \mathbb{E}_{\xi_i^{k,h}} [\nabla f_i(x^{k,h}, \xi_i^{k,h}) - \nabla f_i(x^{k,h})], \mathbb{E}_{\xi_j^{k,h}} [\nabla f_j(x^{k,h}, \xi_j^{k,h}) - \nabla f_j(x^{k,h})] \right\rangle \\
& \stackrel{\text{As. 2.4}}{\leq} \frac{1}{M} \sigma^2,
\end{aligned}$$

1828 where (i) is correct, since $\xi_i^{k,h}$ and $\xi_j^{k,h}$ are independent. Substituting this estimate into equation 23,
1829 we have

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

$$\begin{aligned}
& + 2\gamma^2 (1 - \theta) \|\nabla f(x^{k,h})\|^2 + \gamma^2 \theta \|g^{k-1,H^{k-1}}\|^2 \\
& + \frac{2\gamma^2(1 - \theta)\sigma^2}{M}.
\end{aligned} \tag{24}$$

Let us estimate scalar products separately.

$$\begin{aligned}
-2\gamma\theta \langle \tilde{x}^{k,h} - x^*, g^{k-1,H^{k-1}} \rangle & \stackrel{(Fen)}{\leq} \theta \|\tilde{x}^{k,h} - x^*\|^2 + \gamma^2 \theta \|g^{k-1,H^{k-1}}\|^2, \\
-2\gamma(1 - \theta) \langle \tilde{x}^{k,h} - x^{k,h}, \nabla f(x^{k,h}) \rangle & \stackrel{(Fen)}{\leq} (1 - \theta) \|\tilde{x}^{k,h} - x^{k,h}\|^2 \\
& + \gamma^2(1 - \theta) \|\nabla f(x^{k,h})\|^2 \\
& \stackrel{(20)}{=} \gamma^2(1 - \theta) \|g^{k,h}\|^2 + \gamma^2(1 - \theta) \|\nabla f(x^{k,h})\|^2, \\
-2\gamma(1 - \theta) \langle x^{k,h} - x^*, \nabla f(x^{k,h}) \rangle & \stackrel{\text{As. 2.2(b)}}{\leq} -\gamma\mu(1 - \theta) \|x^{k,h} - x^*\|^2 \\
& - 2\gamma(1 - \theta) [f(x^{k,h}) - f(x^*)] \\
& \stackrel{(CS)}{\leq} -\frac{\gamma\mu(1 - \theta)}{2} \|\tilde{x}^{k,h} - x^*\|^2 \\
& + \gamma\mu(1 - \theta) \|x^{k,h} - \tilde{x}^{k,h}\|^2 \\
& - 2\gamma(1 - \theta) [f(x^{k,h}) - f(x^*)] \\
& \stackrel{(20)}{=} -\frac{\gamma\mu(1 - \theta)}{2} \|\tilde{x}^{k,h} - x^*\|^2 \\
& + \gamma^3\mu(1 - \theta) \|g^{k,h}\|^2 \\
& - 2\gamma(1 - \theta) [f(x^{k,h}) - f(x^*)].
\end{aligned}$$

Substituting these estimates into equation 24, we obtain

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

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 this way, equation 25 transforms to

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

Next we estimate

1890

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

1893 and combine with equation 25:

1894

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

1902 By choosing $\gamma \leqslant \frac{1}{6L}$ we can simplify as
1903

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

1910 Now we put $h = H^k - 1$ and take additional expectations to obtain
1911

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

1922 We take expectation with respect to H^{k-1} and H^k , and apply Lemma C.1:
1923

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

1942 We rearrange terms and use that H^k and $\{\xi_m^{k-1,h}\}_{h=0}^{H^{k-1}-1}$ are independent stochastic values:
1943

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

We use Lemma 3.1 to estimate $\left\| g^{k,H^k} \right\|^2$ and $\left\| g^{k-1,H^{k-1}} \right\|^2$. We obtain

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

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

$$\begin{aligned}
& \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1-1}}} \left\| g^{k-1,H^{k-1}} \right\|^2} \\
& \leq \frac{24(1-\theta)^2\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \left\| \nabla f(x^{k-1,H^{k-1}}) \right\|^2 + \frac{48(1-\theta)^2\alpha\delta_2}{p^2} + \frac{24(1-\theta)^2\alpha\sigma^2}{Mp^2} \\
& \stackrel{(CS)}{\leq} \frac{48(1-\theta)\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \left\| \nabla f(x^{k-1,H^{k-1}}) - \nabla f(\tilde{x}^{k-1,H^{k-1}}) \right\|^2 \\
& \quad + \frac{48(1-\theta)^2\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \left\| \nabla f(\tilde{x}^{k-1,H^{k-1}}) \right\|^2 + \frac{48(1-\theta)^2\alpha\delta_2}{p^2} + \frac{24(1-\theta)^2\alpha\sigma^2}{Mp^2} \\
& \stackrel{\text{As. 2.1}}{\leq} \frac{48L^2(1-\theta)^2\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \left\| x^{k-1,H^{k-1}} - \tilde{x}^{k-1,H^{k-1}} \right\|^2 \\
& \quad + \frac{48(1-\theta)^2\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \left\| \nabla f(x^{k,0}) \right\|^2 + \frac{48(1-\theta)^2\alpha\delta_2}{p^2} + \frac{24(1-\theta)^2\alpha\sigma^2}{Mp^2} \\
& \stackrel{(20)}{=} \frac{48\gamma^2L^2(1-\theta)^2\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \left\| g^{k-1,H^{k-1}} \right\|^2 \\
& \quad + \frac{48(1-\theta)^2\alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \left\| \nabla f(x^{k,0}) \right\|^2 + \frac{48(1-\theta)^2\alpha\delta_2}{p^2} + \frac{24(1-\theta)^2\alpha\sigma^2}{Mp^2}.
\end{aligned}$$

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

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

Applying equation Lip to equation 28, equation 29 and substituting it to equation 27, we get

$$\begin{aligned}
& 1998 \\
& 1999 \\
& 2000 \quad p \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0} \dots \mathbb{E}_{\xi_m^{k,H^{k-1}}} \left\| \tilde{x}^{k,H^k} - x^* \right\|^2} \\
& 2001 \quad \leq p \left(1 - \frac{\gamma\mu}{8} \right) \mathbb{E}_{\xi_m^{k-1,0} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}}} \left\| \tilde{x}^{k,0} - x^* \right\|^2} \\
& 2002 \quad - \gamma(1-p)(1-\theta) \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0} \dots \mathbb{E}_{\xi_m^{k,H^{k-1}}} \left[f(x^{k,H^k}) - f(x^*) \right]} \\
& 2003 \quad - \gamma p(1-\theta) \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}}} \left[f(x^{k,0}) - f(x^*) \right]} \\
& 2004 \quad + \frac{96\gamma^2 L(1-p)(1-\theta)^3 \alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0} \dots \mathbb{E}_{\xi_m^{k-1,H-1}} \mathbb{E}_{H^k} \left[f(x^{k,H^k}) - f(x^*) \right]} \\
& 2005 \quad + \frac{96\gamma^2(1-p)(1-\theta)^3 \alpha \delta_2}{p^2} + \frac{48\gamma^2(1-p)(1-\theta)^3 \alpha \sigma^2}{Mp^2} \\
& 2006 \quad + \frac{384\gamma^2 L \theta(1-\theta)^2 \alpha(\delta_1+1)}{p^2} \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0} \dots \mathbb{E}_{\xi_m^{k-1,H-1}} \left[f(x^{k,0}) - f(x^*) \right]} \\
& 2007 \quad + \frac{192\gamma^2 \theta(1-\theta)^2 \alpha \delta_2}{p^2} + \frac{96\gamma^2 \theta(1-\theta)^2 \alpha \sigma^2}{Mp^2} + \frac{2\gamma^2(1-\theta)\sigma^2}{M}.
\end{aligned}$$

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

$$\begin{aligned}
& 2016 \\
& 2017 \\
& 2018 \quad p \mathbb{E} \left\| \tilde{x}^{k,H^k} - x^* \right\|^2 \leq p \left(1 - \frac{\gamma\mu}{8} \right) \mathbb{E} \left\| \tilde{x}^{k,0} - x^* \right\|^2 \\
& 2019 \quad - \gamma(1-p)(1-\theta) \left(1 - \frac{96\gamma L(1-\theta)^2 \alpha(\delta_1+1)}{p^2} \right) \cdot \\
& 2020 \quad \cdot \mathbb{E} \left[f(x^{k,H^k}) - f(x^*) \right] \\
& 2021 \quad - \gamma p(1-\theta) \left(1 - \frac{384\gamma L \theta(1-\theta) \alpha(\delta_1+1)}{p^3} \right) \mathbb{E} \left[f(x^{k,0}) - f(x^*) \right] \\
& 2022 \quad + \frac{96\gamma^2(1-p)(1-\theta)^3 \alpha \delta_2}{p^2} + \frac{192\gamma^2 \theta(1-\theta)^2 \alpha \delta_2}{p^2} \\
& 2023 \quad + \frac{48\gamma^2(1-p)(1-\theta)^3 \alpha \sigma^2}{Mp^2} + \frac{96\gamma^2 \theta(1-\theta)^2 \alpha \sigma^2}{Mp^2} \\
& 2024 \quad + \frac{2\gamma^2(1-\theta)\sigma^2}{M}.
\end{aligned}$$

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 choice of θ and γ , consequently this choice is correct. In that way, we obtain

$$\begin{aligned}
& 2025 \\
& 2026 \quad \mathbb{E} \left\| \tilde{x}^{k,H^k} - x^* \right\|^2 \leq \left(1 - \frac{\gamma\mu}{8} \right) \mathbb{E} \left\| \tilde{x}^{k,0} - x^* \right\|^2 + \frac{96\gamma^2 \alpha \delta_2}{p^3} + \frac{48\gamma^3 \mu \alpha \delta_2}{p^2} \\
& 2027 \quad + \frac{48\gamma^2 \alpha \sigma^2}{Mp^3} + \frac{24\gamma^3 \mu \alpha \sigma^2}{Mp^2} + \frac{2\gamma^2 \sigma^2}{Mp}.
\end{aligned}$$

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,

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

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

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

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

□

Corollary D.7 (Corollary 3.6). *Under conditions of Theorem 3.5 Algorithm 1 with fixed rules $\widehat{\mathcal{R}} \equiv \widetilde{\mathcal{R}} \equiv \mathcal{R}$ needs*

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

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

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

Proof. Using the result of Theorem 3.5, we choose

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

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

Corollary D.8. *Under conditions of Theorem 3.5 Algorithm 1 needs*

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

number of devices communications 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 k -th iteration in h -th epoch.

Proof. Using the result of Theorem 3.5, we choose

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

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 $\varepsilon^2 = \mathbb{E} \|x^{K,0} - x^*\|^2$. Since the length of the epoch $H \in \text{Geom}(p)$, Algorithm 1 requires $\widetilde{\mathcal{O}}\left(\frac{L\alpha\delta_1}{\mu p^3} \log\left(\frac{1}{\varepsilon}\right) + \frac{\alpha\delta_2}{\mu^2 p^4 \varepsilon} + \frac{\alpha\sigma^2}{\mu^2 M p^4 \varepsilon}\right)$ communication rounds. Next we mention that at each communication round we communicate with $C^{k,h}$ devices, thus, number of communications is $\widetilde{\mathcal{O}}\left(\max_{k,h} C^{k,h} \left(\frac{L\alpha\delta_1}{\mu p^3} \log\left(\frac{1}{\varepsilon}\right) + \frac{\alpha\delta_2}{\mu^2 p^4 \varepsilon} + \frac{\alpha\sigma^2}{\mu^2 M p^4 \varepsilon}\right)\right)$. Taking, $p = \frac{\min_{k,h} C^{k,h}}{M}$, we have the result of the

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 D.9. Considering fixed rules $\widehat{\mathcal{R}} \equiv \widetilde{\mathcal{R}} \equiv \mathcal{R}$, we have $\widetilde{\mathcal{O}}\left(M\left(\frac{M}{C}\right)^2\left(\frac{L}{\mu}\delta_1 \log\left(\frac{1}{\varepsilon}\right) + \frac{M}{C}\frac{\delta_2}{\mu^2\varepsilon} + \frac{\sigma^2}{\mu^2C\varepsilon}\right)\right)$ and $\widetilde{\mathcal{O}}\left(M^2\left(\frac{M}{C}\right)^2\left(\frac{L}{\mu}\delta_1 \log\left(\frac{1}{\varepsilon}\right) + \frac{M}{C}\frac{\delta_2}{\mu^2\varepsilon} + \frac{\sigma^2}{\mu^2C\varepsilon}\right)\right)$ 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 E.1. Suppose Assumptions 2.3, 2.4 hold. Then for Algorithm 2 it implies that

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

Proof. Let us start with the following estimate:

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

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

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

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

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

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

In that way,

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

We obtained an estimate for the second term in Lemma D.1 in equation 7:

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

Moreover, in equation 6 we found out

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

Combining this estimates with equation 31,

$$\begin{aligned} & \mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} \left\| \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \left(\frac{1}{M} - \hat{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\ & \leq \frac{2\alpha}{\min_{1 \leq m \leq M} q_m M} \sum_{m=1}^M \mathbb{E}_{\xi_m^{k,h}} \left\| \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\ & \quad + 8\alpha(\delta_1 + 1) \left\| \nabla f(x^{k,h}) \right\|^2 + 8\alpha\delta_2 + \frac{4\alpha\sigma^2}{M} \\ & \stackrel{(CS)}{\leq} \frac{4\alpha}{\min_{1 \leq m \leq M} q_m M} \sum_{m=1}^M \mathbb{E}_{\xi_m^{k,h}} \left\| \nabla f_m(x^{k,h}) \right\|^2 \\ & \quad + \frac{4\alpha}{\min_{1 \leq m \leq M} q_m M} \sum_{m=1}^M \mathbb{E}_{\xi_m^{k,h}} \left\| \nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h}) \right\|^2 \\ & \quad + 8\alpha(\delta_1 + 1) \left\| \nabla f(x^{k,h}) \right\|^2 + 8\alpha\delta_2 + \frac{4\alpha\sigma^2}{M} \\ & \stackrel{\text{As. 2.4}}{\leq} \frac{4\alpha}{\min_{1 \leq m \leq M} q_m M} \sum_{m=1}^M \left\| \nabla f_m(x^{k,h}) \right\|^2 + \frac{4\alpha\sigma^2}{\min_{1 \leq m \leq M} q_m} \\ & \quad + 8\alpha(\delta_1 + 1) \left\| \nabla f(x^{k,h}) \right\|^2 + 8\alpha\delta_2 + \frac{4\alpha\sigma^2}{M} \end{aligned}$$

$$\begin{aligned}
& \stackrel{(CS)}{\leq} \frac{8\alpha}{\min_{1 \leq m \leq M} q_m M} \sum_{m=1}^M \|\nabla f(x^{k,h})\|^2 + \frac{8\alpha}{\min_{1 \leq m \leq M} q_m M} \sum_{m=1}^M \|\nabla f_m(x^{k,h}) - \nabla f(x^{k,h})\|^2 \\
& + 8\alpha(\delta_1 + 1) \|\nabla f(x^{k,h})\|^2 + 8\alpha\delta_2 + \frac{8\alpha\sigma^2}{\min_{1 \leq m \leq M} q_m} \\
& \stackrel{\text{As. 2.3}}{\leq} \frac{8\alpha(\delta_1 + 1)}{\min_{1 \leq m \leq M} q_m} \|\nabla f(x^{k,h})\|^2 + 8\alpha(\delta_1 + 1) \|\nabla f(x^{k,h})\|^2 + \frac{8\alpha\delta_2}{\min_{1 \leq m \leq M} q_m} \\
& + 8\alpha\delta_2 + \frac{8\alpha\sigma^2}{\min_{1 \leq m \leq M} q_m} \\
& \leq \frac{16\alpha(\delta_1 + 1)}{\min_{1 \leq m \leq M} q_m} \|\nabla f(x^{k,h})\|^2 + \frac{16\alpha\delta_2}{\min_{1 \leq m \leq M} q_m} + \frac{8\alpha\sigma^2}{\min_{1 \leq m \leq M} q_m}.
\end{aligned}$$

Substituting this estimate into equation 30, we have

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

Enrolling a recursion, we get

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

Next, choosing $c = \frac{p}{2}$, taking exception on H^k and applying equation 9, equation 10, equation 11 from Lemma D.1, we obtain the result of the lemma:

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

□

2261
2262
2263
2264
2265
2266
2267

2268 E.1 PROOF FOR NON-CONVEX SETTING

2269
Theorem E.2. Suppose Assumptions 2.1, 2.2(a), 2.3, 2.4 hold. Then for Algorithm 2 with $\theta \leq \frac{\gamma L p^2}{2}$
2270 and $\gamma \leq \frac{p \min_{1 \leq m \leq M} q_m}{768 L \alpha (\delta_1 + 1)}$ it implies that

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

2281 *Proof.* We start with the definition of virtual sequence:

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

2285 It is followed by

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

2298 Assumption 2.1 implies

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

2310 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
2311 $x^{k,h}, \tilde{x}^{k,h}, \xi_m^{k,h}, g^{k-1, H^{k-1}}$, we take the expectation and obtain

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

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

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

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,

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

Let us consider separately the following term:

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

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

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

Combining with equation 36,

$$\mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2$$

$$\begin{aligned}
& \leq \frac{2}{M \min_{1 \leq m \leq M} q_m} \sum_{m=1}^M \mathbb{E}_{\xi_m^{k,h}} \|\nabla f_m(x^{k,h}, \xi_m^{k,h})\|^2 + 2\mathbb{E}_{\xi_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\
& \stackrel{(CS)}{\leq} \frac{4}{M \min_{1 \leq m \leq M} q_m} \sum_{m=1}^M \mathbb{E}_{\xi_m^{k,h}} \|\nabla f_m(x^{k,h}, \xi_m^{k,h})\|^2 \\
& \stackrel{(CS)}{\leq} \frac{8}{M \min_{1 \leq m \leq M} q_m} \sum_{m=1}^M \|\nabla f_m(x^{k,h})\|^2 \\
& \quad + \frac{8}{M \min_{1 \leq m \leq M} q_m} \sum_{m=1}^M \mathbb{E}_{\xi_m^{k,h}} \|\nabla f_m(x^{k,h}, \xi_m^{k,h}) - \nabla f_m(x^{k,h})\|^2 \\
& \stackrel{\text{As. 2.4}}{\leq} \frac{8}{M \min_{1 \leq m \leq M} q_m} \sum_{m=1}^M \|\nabla f_m(x^{k,h})\|^2 + \frac{8\sigma^2}{\min_{1 \leq m \leq M} q_m} \\
& \stackrel{(CS)}{\leq} \frac{16}{M \min_{1 \leq m \leq M} q_m} \sum_{m=1}^M \|\nabla f(x^{k,h})\|^2 \\
& \quad + \frac{16}{M \min_{1 \leq m \leq M} q_m} \sum_{m=1}^M \|\nabla f_m(x^{k,h}) - \nabla f(x^{k,h})\|^2 + \frac{8\sigma^2}{\min_{1 \leq m \leq M} q_m} \\
& \stackrel{\text{As. 2.3}}{\leq} \frac{16(\delta_1 + 1)}{\min_{1 \leq m \leq M} q_m} \|\nabla f(x^{k,h})\|^2 + \frac{16\delta_2}{\min_{1 \leq m \leq M} q_m} + \frac{8\sigma^2}{\min_{1 \leq m \leq M} q_m}. \tag{37}
\end{aligned}$$

We substitute this estimate into equation 35 to obtain

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

Let us estimate the scalar products separately.

$$\begin{aligned}
-\gamma(1-\theta) \langle \nabla f(\tilde{x}^{k,h}), \nabla f(x^{k,h}) \rangle & = -\frac{\gamma(1-\theta)}{2} \|\nabla f(\tilde{x}^{k,h})\|^2 - \frac{\gamma(1-\theta)}{2} \|\nabla f(x^{k,h})\|^2 \\
& \quad + \frac{\gamma(1-\theta)}{2} \|\nabla f(\tilde{x}^{k,h}) - \nabla f(x^{k,h})\|^2 \\
& \stackrel{\text{As. 2.1}}{\leq} -\frac{\gamma(1-\theta)}{2} \|\nabla f(\tilde{x}^{k,h})\|^2 - \frac{\gamma(1-\theta)}{2} \|\nabla f(x^{k,h})\|^2 \\
& \quad + \frac{\gamma L^2(1-\theta)}{2} \|\tilde{x}^{k,h} - x^{k,h}\|^2 \\
& \stackrel{(12)}{=} -\frac{\gamma(1-\theta)}{2} \|\nabla f(\tilde{x}^{k,h})\|^2 - \frac{\gamma(1-\theta)}{2} \|\nabla f(x^{k,h})\|^2 \\
& \quad + \frac{\gamma^3 L^2(1-\theta)}{2} \|g^{k,h}\|^2, \\
-\gamma\theta \left\langle \nabla f(\tilde{x}^{k,h}), g^{k-1, H^{k-1}} \right\rangle & \stackrel{(Fen)}{\leq} \frac{\gamma\theta}{2} \|\nabla f(\tilde{x}^{k,h})\|^2 + \frac{\gamma\theta}{2} \|g^{k-1, H^{k-1}}\|^2.
\end{aligned}$$

2430 Combining it with equation 38,

$$\begin{aligned}
 \mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} [f(\tilde{x}^{k,h+1})] &\leq f(\tilde{x}^{k,h}) - \frac{\gamma(1-\theta)}{2} \left(1 - \frac{16\gamma L(\delta_1+1)}{\min_{1 \leq m \leq M} q_m} \right) \|\nabla f(x^{k,h})\|^2 \\
 &\quad - \frac{\gamma(1-2\theta)}{2} \|\nabla f(\tilde{x}^{k,h})\|^2 + \frac{\gamma^3 L^2(1-\theta)}{2} \|g^{k,h}\|^2 \\
 &\quad + \frac{\gamma\theta(\gamma L+1)}{2} \|g^{k-1,H^{k-1}}\|^2 + \frac{8\gamma^2 L(1-\theta)\delta_2}{\min_{1 \leq m \leq M} q_m} + \frac{4\gamma^2 L(1-\theta)\sigma^2}{\min_{1 \leq m \leq M} q_m}.
 \end{aligned}$$

2441 Choosing $\gamma \leq \frac{\min_{1 \leq m \leq M} q_m}{32L(\delta_1+1)}$ and $\theta \leq \frac{1}{2}$,

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

2448 Now we put $h = H^k - 1$ and take additional expectations.

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

2463 We take expectation with respect to H^{k-1} and H^k , and apply Lemma C.1:

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

2484 We use Lemma E.1 to estimate $\|g^{k,H^k}\|^2$ and $\|g^{k-1,H^{k-1}}\|^2$. We have
 2485

$$\begin{aligned} 2486 \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k,0}} \mathbb{E}_{\eta_m^{k,0}} \dots \mathbb{E}_{\xi_m^{k,H^{k-1}}} \mathbb{E}_{\eta_m^{k,H^{k-1}}} \|g^{k,H^k}\|^2 &\leq \frac{96(1-\theta)^2\alpha(\delta_1+1)}{p^2 \min_{1 \leq m \leq M} q_m} \mathbb{E}_{H^k} \|\nabla f(x^{k,H^k})\|^2 \\ 2487 &+ \frac{192(1-\theta)^2\alpha\delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{96(1-\theta)^2\alpha\sigma^2}{p^2 \min_{1 \leq m \leq M} q_m}. \end{aligned} \quad (40)$$

2492 Next, analogously to equation 19, we choose $\gamma \leq \frac{p \cdot \sqrt{\min_{1 \leq m \leq M} q_m}}{384L\sqrt{\alpha}\sqrt{\delta_1+1}}$ and obtain
 2493

$$\begin{aligned} 2494 \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \mathbb{E}_{\eta_m^{k-1,H^{k-1}-1}} \|g^{k-1,H^{k-1}}\|^2 \\ 2495 &\leq \frac{384(1-\theta)^2\alpha(\delta_1+1)}{p^2 \min_{1 \leq m \leq M} q_m} \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^{k-1}-1}} \mathbb{E}_{\eta_m^{k-1,H^{k-1}-1}} \|\nabla f(x^{k,0})\|^2 \\ 2496 &+ \frac{384(1-\theta)^2\alpha\delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{192(1-\theta)^2\alpha\sigma^2}{p^2 \min_{1 \leq m \leq M} q_m}. \end{aligned} \quad (41)$$

2503 We combine equation 40 and equation 41 with equation 39 and use that H^{k-1} with $\{\xi_m^{k-1,h}\}_{h=0}^{H^{k-1}-1}$
 2504 and H^k with $\{\eta_m^{k-1,h}\}_{h=0}^{H^{k-1}-1}$ are independent stochastic values. Moreover we take full expectation:
 2505

$$\begin{aligned} 2506 p\mathbb{E}[f(\tilde{x}^{k,H^k})] &\leq p\mathbb{E}[f(\tilde{x}^{k,0})] \\ 2507 &- \frac{\gamma(1-\theta)p}{4} \mathbb{E}\|\nabla f(x^{k,0})\|^2 - \frac{\gamma(1-\theta)(1-p)}{4} \mathbb{E}\|\nabla f(x^{k,H^k})\|^2 \\ 2508 &+ \frac{48\gamma^3L^2(1-\theta)^3(1-p)\alpha(\delta_1+1)}{p^2 \min_{1 \leq m \leq M} q_m} \mathbb{E}\|\nabla f(x^{k,H^k})\|^2 \\ 2509 &+ \frac{192\gamma\theta(1-\theta)^2\alpha(\delta_1+1)}{p^2 \min_{1 \leq m \leq M} q_m} \mathbb{E}\|\nabla f(x^{k,0})\|^2 \\ 2510 &+ \frac{96\gamma^3L^2(1-\theta)^3(1-p)\alpha\delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{384\gamma\theta(1-\theta)^2\alpha\delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{8\gamma^2L(1-\theta)\delta_2}{\min_{1 \leq m \leq M} q_m} \\ 2511 &+ \frac{48\gamma^3L^2(1-\theta)^3(1-p)\alpha\sigma^2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{192\gamma\theta(1-\theta)^2\alpha\sigma^2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{4\gamma^2L(1-\theta)\sigma^2}{\min_{1 \leq m \leq M} q_m} \\ 2512 &= p\mathbb{E}[f(\tilde{x}^{k,0})] \\ 2513 &- \frac{\gamma(1-\theta)(1-p)}{4} \left(1 - \frac{192\gamma^2L^2(1-\theta)^2\alpha(\delta_1+1)}{p^2 \min_{1 \leq m \leq M} q_m}\right) \mathbb{E}\|\nabla f(x^{k,H^k})\|^2 \\ 2514 &- \frac{\gamma(1-\theta)p}{4} \left(1 - \frac{384\theta(1-\theta)^2\alpha(\delta_1+1)}{p^3 \min_{1 \leq m \leq M} q_m}\right) \mathbb{E}\|\nabla f(x^{k,0})\|^2 \\ 2515 &+ \frac{96\gamma^3L^2(1-\theta)^3(1-p)\alpha\delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{768\gamma\theta(1-\theta)^2\alpha\delta_2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{8\gamma^2L(1-\theta)\delta_2}{\min_{1 \leq m \leq M} q_m} \\ 2516 &+ \frac{48\gamma^3L^2(1-\theta)^3(1-p)\alpha\sigma^2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{192\gamma\theta(1-\theta)^2\alpha\sigma^2}{p^2 \min_{1 \leq m \leq M} q_m} + \frac{4\gamma^2L(1-\theta)\sigma^2}{\min_{1 \leq m \leq M} q_m}. \end{aligned}$$

2536 We choose $\theta \leq \frac{\gamma L p^2}{2}$ $\gamma \leq \frac{p \min_{1 \leq m \leq M} q_m}{768L\alpha(\delta+1)}$. In that way,
 2537

2538

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

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,

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

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

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

□

Corollary E.3 (Corollary 3.8). *Under conditions of Theorem E.2 Algorithm 2 with fixed rules $\hat{\mathcal{R}} \equiv \tilde{\mathcal{R}} \equiv \mathcal{R}$ needs*

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

2592 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
 2593 of devices participating in each epoch.
 2594

2596 *Proof.* Proof is analogous to the proof of Corollary D.3. \square
 2597

2598 **Corollary E.4.** *Under conditions of Theorem E.2 Algorithm 2 needs*

$$2599 \mathcal{O} \left(\frac{M}{\min_{k,h} C^{k,h}} \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{\Delta L \alpha \delta_1}{\varepsilon^2} + \frac{\Delta L \alpha \delta_2}{\varepsilon^4} + \frac{\Delta L \alpha \sigma^2}{\varepsilon^4} \right) \right) \text{ epochs and}$$

$$2600 \mathcal{O} \left(M \left(\frac{M}{\min_{k,h} C^{k,h}} \right)^2 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{\Delta L \alpha \delta_1}{\varepsilon^2} + \frac{\Delta L \alpha \delta_2}{\varepsilon^4} + \frac{\Delta L \alpha \sigma^2}{\varepsilon^4} \right) \right)$$

2601 number of devices communications to reach ε -accuracy, where $\varepsilon^2 = \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(x^{k,0})\|^2$, $\Delta =$
 2602 $f(x^{0,0}) - f(x^*)$ and $C^{k,h}$ is the number of devices participating in k -th iteration in h -th epoch.
 2603

2604 *Proof.* Proof is analogous to the proof of Corollary D.4. \square
 2605

2606 **Remark E.5.** Considering fixed rules $\widehat{\mathcal{R}} \equiv \widetilde{\mathcal{R}} \equiv \mathcal{R}$,
 2607 we have $\mathcal{O} \left(M \frac{M}{C} \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{\Delta L \delta_1}{\varepsilon^2} + \frac{\Delta L \delta_2}{\varepsilon^4} + \frac{\Delta L \sigma^2}{M \varepsilon^4} \right) \right)$
 2608 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-
 2609 ularizing parameter $\alpha = 1$ and $\alpha = M$ respectively. Considering various rules, best case with
 2610 regularizing coefficient $\alpha = 1$ gives us $\mathcal{O} \left(M \left(\frac{M}{\min_{k,h} C^{k,h}} \right)^2 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{\Delta L \delta_1}{\varepsilon^2} + \frac{\Delta L \delta_2}{\varepsilon^4} + \frac{\Delta L \sigma^2}{M \varepsilon^4} \right) \right)$
 2611 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-
 2612 ber of devices communications.
 2613

2614 E.2 PROOF FOR STRONGLY-CONVEX SETTING

2615 **Theorem E.6.** *Suppose Assumptions 2.1, 2.2(b), 2.3, 2.4 hold. Then for Algorithm 2 with $\theta \leq \frac{p\gamma\mu}{4}$
 2616 and $\gamma \leq \frac{p^2 \min_{1 \leq m \leq M} q_m}{384L\alpha(\delta_1+1)}$ it implies that*

$$2617 \mathbb{E} \|x^{K,0} - x^*\|^2 \leq \left(1 - \frac{\gamma\mu}{8}\right)^K \mathbb{E} \|x^{0,0} - x^*\|^2 + \frac{2368\gamma\alpha}{\mu p^3} \min_{1 \leq m \leq M} q_m (2\delta_2 + \sigma^2).$$

2618 *Proof.* We start with the definition of virtual sequence:
 2619

$$2620 \widetilde{x}^{k,h} = x^{k,h} - \gamma \sum_{m=1}^M g_m^{k,h} = x^{k,h} - \gamma g^{k,h}. \quad (42)$$

2621 It is followed by

$$2622 \begin{aligned} \widetilde{x}^{k,h+1} &= x^{k,h+1} - \gamma \sum_{m=1}^M g_m^{k,h+1} \\ 2623 &= x^{k,h} - \gamma \left[(1 - \theta) \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \hat{\pi}_m^{k,h} \nabla f_m(x^{k,h}, \xi_m^{k,h}) + \theta g^{k-1, H^{k-1}} \right] \\ 2624 &\quad - \gamma \sum_{m=1}^M g_m^{k,h} - \gamma (1 - \theta) \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \left(\frac{1}{M} - \hat{\pi}_m^{k,h} \right) \nabla f_m(x^{k,h}, \xi_m^{k,h}) \end{aligned}$$

$$2646 \quad = \quad \tilde{x}^{k,h} - \gamma \left[(1-\theta) \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) + \theta g^{k-1, H^{k-1}} \right]. \quad (43)$$

2649 Next, we use this to write a descent:

$$\begin{aligned} 2650 \quad \|\tilde{x}^{k,h+1} - x^*\|^2 &= \|\tilde{x}^{k,h} - x^*\|^2 + 2 \langle \tilde{x}^{k,h} - x^*, \tilde{x}^{k,h+1} - \tilde{x}^{k,h} \rangle + \|\tilde{x}^{k,h+1} - \tilde{x}^{k,h}\|^2 \\ 2651 \quad &\stackrel{(43)}{=} \|\tilde{x}^{k,h} - x^*\|^2 - 2\gamma\theta \langle \tilde{x}^{k,h} - x^*, g^{k-1, H^{k-1}} \rangle \\ 2652 \quad &\quad - 2\gamma(1-\theta) \left\langle \tilde{x}^{k,h} - x^*, \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ 2653 \quad &\quad + \gamma^2 \left\| \theta g^{k-1, H^{k-1}} + (1-\theta) \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2 \\ 2654 \quad &\stackrel{(Jen)}{\leqslant} \|\tilde{x}^{k,h} - x^*\|^2 - 2\gamma\theta \langle \tilde{x}^{k,h} - x^*, g^{k-1, H^{k-1}} \rangle \\ 2655 \quad &\quad - 2\gamma(1-\theta) \left\langle \tilde{x}^{k,h} - x^{k,h}, \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ 2656 \quad &\quad - 2\gamma(1-\theta) \left\langle x^{k,h} - x^*, \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ 2657 \quad &\quad + \gamma^2 \theta \left\| g^{k-1, H^{k-1}} \right\|^2 + \gamma^2(1-\theta) \left\| \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2. \end{aligned}$$

2658 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 $x^{k,h}, \tilde{x}^{k,h}, \xi_m^{k,h}, g^{k-1, H^{k-1}}$, we take the expectation and obtain

$$\begin{aligned} 2659 \quad \mathbb{E}_{\eta_m^{k,h}} \|\tilde{x}^{k,h+1} - x^*\|^2 &\leq \|\tilde{x}^{k,h} - x^*\|^2 - 2\gamma\theta \langle \tilde{x}^{k,h} - x^*, g^{k-1, H^{k-1}} \rangle \\ 2660 \quad &\quad - 2\gamma(1-\theta) \left\langle \tilde{x}^{k,h} - x^{k,h}, \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ 2661 \quad &\quad - 2\gamma(1-\theta) \left\langle x^{k,h} - x^*, \frac{1}{M} \sum_{m=1}^M \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\rangle \\ 2662 \quad &\quad + \gamma^2 \theta \left\| g^{k-1, H^{k-1}} \right\|^2 \\ 2663 \quad &\quad + \gamma^2(1-\theta) \mathbb{E}_{\eta_m^{k,h}} \left\| \frac{1}{M} \sum_{m=1}^M \frac{\eta_m^{k,h}}{q_m} \nabla f_m(x^{k,h}, \xi_m^{k,h}) \right\|^2. \end{aligned}$$

2664 Now we take the expectation over $\xi_m^{k,h}$. Mention that

$$\begin{aligned} 2665 \quad \tilde{x}^{k,h} &\stackrel{(42)}{=} x^{k,h} - \gamma g^{k,h} \\ 2666 \quad &\stackrel{\text{Line 12}}{=} x^{k,h} - \gamma \left(g^{k,h-1} + (1-\theta) \sum_{m=1}^M \frac{\eta_m^{k,h-1}}{q_m} \left(\frac{1}{M} - \hat{\pi}_m^{k,h-1} \right) \nabla f(x^{k,h-1}, \xi_m^{k,h-1}) \right). \end{aligned}$$

2667 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,

$$\begin{aligned} 2668 \quad \mathbb{E}_{\xi_m^{k,h}} \mathbb{E}_{\eta_m^{k,h}} \|\tilde{x}^{k,h+1} - x^*\|^2 &\leq \|\tilde{x}^{k,h} - x^*\|^2 - 2\gamma\theta \langle \tilde{x}^{k,h} - x^*, g^{k-1, H^{k-1}} \rangle \\ 2669 \quad &\quad - 2\gamma(1-\theta) \langle \tilde{x}^{k,h} - x^{k,h}, \nabla f(x^{k,h}) \rangle \\ 2670 \quad &\quad - 2\gamma(1-\theta) \langle x^{k,h} - x^*, \nabla f(x^{k,h}) \rangle \\ 2671 \quad &\quad + \gamma^2 \theta \left\| g^{k-1, H^{k-1}} \right\|^2 \end{aligned}$$

$$2700 \quad + \gamma^2(1-\theta)\mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\left\|\frac{1}{M}\sum_{m=1}^M\frac{\eta_m^{k,h}}{q_m}\nabla f_m(x^{k,h},\xi_m^{k,h})\right\|^2. \quad (44)$$

2703 Recall we estimated the last term in Theorem E.2 in equation 37:

$$2705 \quad \mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\left\|\frac{1}{M}\sum_{m=1}^M\frac{\eta_m^{k,h}}{q_m}\nabla f_m(x^{k,h},\xi_m^{k,h})\right\|^2 \leq \frac{16(\delta_1+1)}{\min_{1\leq m\leq M}q_m}\|\nabla f(x^{k,h})\|^2 + \frac{16\delta_2}{\min_{1\leq m\leq M}q_m} \\ 2706 \quad + \frac{8\sigma^2}{\min_{1\leq m\leq M}q_m}. \quad (45)$$

2711 Now let us estimate scalar products separately.

$$2713 \quad -2\gamma\theta\left\langle\tilde{x}^{k,h}-x^*,g^{k-1,H^{k-1}}\right\rangle \stackrel{(Fen)}{\leq} \theta\|\tilde{x}^{k,h}-x^*\|^2 + \gamma^2\theta\|g^{k-1,H^{k-1}}\|^2, \\ 2714 \quad -2\gamma(1-\theta)\left\langle\tilde{x}^{k,h}-x^{k,h},\nabla f(x^{k,h})\right\rangle \stackrel{(Fen)}{\leq} (1-\theta)\|\tilde{x}^{k,h}-x^{k,h}\|^2 \\ 2715 \quad + \gamma^2(1-\theta)\|\nabla f(x^{k,h})\|^2 \\ 2716 \quad \stackrel{(20)}{=} \gamma^2(1-\theta)\|g^{k,h}\|^2 + \gamma^2(1-\theta)\|\nabla f(x^{k,h})\|^2, \\ 2717 \quad -2\gamma(1-\theta)\left\langle x^{k,h}-x^*,\nabla f(x^{k,h})\right\rangle \stackrel{\text{As. 2.2(b)}}{\leq} -\gamma\mu(1-\theta)\|x^{k,h}-x^*\|^2 \\ 2718 \quad -2\gamma(1-\theta)[f(x^{k,h})-f(x^*)] \\ 2719 \quad \stackrel{(CS)}{\leq} -\frac{\gamma\mu(1-\theta)}{2}\|\tilde{x}^{k,h}-x^*\|^2 \\ 2720 \quad + \gamma\mu(1-\theta)\|x^{k,h}-\tilde{x}^{k,h}\|^2 \\ 2721 \quad -2\gamma(1-\theta)[f(x^{k,h})-f(x^*)] \\ 2722 \quad \stackrel{(20)}{=} -\frac{\gamma\mu(1-\theta)}{2}\|\tilde{x}^{k,h}-x^*\|^2 \\ 2723 \quad + \gamma^3\mu(1-\theta)\|g^{k,h}\|^2 \\ 2724 \quad -2\gamma(1-\theta)[f(x^{k,h})-f(x^*)].$$

2725 Substituting this estimates and equation 45 into equation 44,

$$2726 \quad \mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\|\tilde{x}^{k,h+1}-x^*\|^2 \leq \left(1-\frac{\gamma\mu(1-\theta)}{2}+\theta\right)\|\tilde{x}^{k,h}-x^*\|^2 + 2\gamma^2\theta\|g^{k-1,H^{k-1}}\|^2 \\ 2727 \quad + \gamma^2(1-\theta)(1+\gamma\mu)\|g^{k,h}\|^2 \\ 2728 \quad + \frac{19\gamma^2(1-\theta)(\delta_1+1)}{\min_{1\leq m\leq M}q_m}\|\nabla f(x^{k,h})\|^2 \\ 2729 \quad -2\gamma(1-\theta)[f(x^{k,h})-f(x^*)] \\ 2730 \quad + \frac{16\gamma^2(1-\theta)\delta_2}{\min_{1\leq m\leq M}q_m} + \frac{8\gamma^2(1-\theta)\sigma^2}{\min_{1\leq m\leq M}q_m}.$$

2731 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
2732 this way,

$$2733 \quad \mathbb{E}_{\xi_m^{k,h}}\mathbb{E}_{\eta_m^{k,h}}\|\tilde{x}^{k,h+1}-x^*\|^2 \leq \left(1-\frac{\gamma\mu}{8}\right)\|\tilde{x}^{k,h}-x^*\|^2 + 2\gamma^2\theta\|g^{k-1,H^{k-1}}\|^2 \\ 2734 \quad + \gamma^2(1-\theta)(1+\gamma\mu)\|g^{k,h}\|^2$$

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

2761 Next, we estimate
2762

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

2766 It implies that
2767

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

2778 Choosing $\gamma \leq \frac{\min_{1 \leq m \leq M} q_m}{38L(\delta_1+1)}$, we can simplify as
2779

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

2788 Now we put $h = H^k - 1$ and take additional expectations to obtain
2789

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

2803 We take expectation with respect to H^{k-1} and H^k , and apply Lemma C.1:
2804

$$\begin{aligned}
& \mathbb{E}_{H^{k-1}} \mathbb{E}_{H^k} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k,H^k-1}} \mathbb{E}_{\eta_m^{k,H^k-1}} \|\tilde{x}^{k,H^k} - x^*\|^2 \\
& \leq p \left(1 - \frac{\gamma\mu}{8}\right) \mathbb{E}_{H^{k-1}} \mathbb{E}_{\xi_m^{k-1,0}} \mathbb{E}_{\eta_m^{k-1,0}} \dots \mathbb{E}_{\xi_m^{k-1,H^k-1}} \mathbb{E}_{\eta_m^{k-1,H^k-1}} \|\tilde{x}^{k,0} - x^*\|^2
\end{aligned}$$

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

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

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

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

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

Now we use equation 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 equation 47 and equation 48 with equation 46 and take full expectation.

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

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

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

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

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,

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

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

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

□

Corollary E.7 (Corollary 3.10). *Under conditions of Theorem E.6 Algorithm 2 with fixed rules $\widehat{\mathcal{R}} \equiv \widehat{\mathcal{R}} \equiv \mathcal{R}$ needs*

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

epochs and

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

number of devices communications

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

Proof. Proof is analogous to the proof of Corollary D.7. □

2916 **Corollary E.8.** *Under conditions of Theorem E.6 Algorithm 2 needs*

$$2917 \tilde{\mathcal{O}} \left(\left(\frac{M}{\min_{k,h} C^{k,h}} \right)^2 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{L}{\mu} \alpha \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{\min_{k,h} C^{k,h}} \frac{\alpha \delta_2}{\mu^2 \varepsilon} + \frac{M}{\min_{k,h} C^{k,h}} \frac{\alpha \sigma^2}{\mu^2 \varepsilon} \right) \right)$$

2921 *epochs or*

$$2922 \tilde{\mathcal{O}} \left(M \left(\frac{M}{\min_{k,h} C^{k,h}} \right)^3 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{L}{\mu} \alpha \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{\min_{k,h} C^{k,h}} \frac{\alpha \delta_2}{\mu^2 \varepsilon} + \frac{M}{\min_{k,h} C^{k,h}} \frac{\alpha \sigma^2}{\mu^2 \varepsilon} \right) \right)$$

2926 *communications*

2927 *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*
 2928 *k -th iteration in h -th epoch.*

2930 *Proof.* Proof is analogous to the proof of Corollary D.8. \square

2932 **Remark E.9.** Considering fixed rules $\widehat{\mathcal{R}} \equiv \widetilde{\mathcal{R}} \equiv \mathcal{R}$,

2933 we have $\widetilde{\mathcal{O}} \left(M \left(\frac{M}{C} \right)^2 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{L}{\mu} \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{C} \frac{\delta_2}{\mu^2 \varepsilon} + \frac{M}{C} \frac{\sigma^2}{\mu^2 \varepsilon} \right) \right)$

2935 and $\widetilde{\mathcal{O}} \left(M^2 \left(\frac{M}{C} \right)^2 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{L}{\mu} \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{C} \frac{\delta_2}{\mu^2 \varepsilon} + \frac{M}{C} \frac{\sigma^2}{\mu^2 \varepsilon} \right) \right)$ number of devices communications with regularizing parameter $\alpha = 1$ and $\alpha = M$ respectively. Considering various rules, best case with regularizing coefficient $\alpha = 1$ gives us
 2939 $\widetilde{\mathcal{O}} \left(M \left(\frac{M}{\min_{k,h} C^{k,h}} \right)^3 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{L}{\mu} \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{\min_{k,h} C^{k,h}} \frac{\delta_2}{\mu^2 \varepsilon} + \frac{M}{\min_{k,h} C^{k,h}} \frac{\sigma^2}{\mu^2 \varepsilon} \right) \right)$ and worst case $\alpha =$
 2941 M gives us $\widetilde{\mathcal{O}} \left(M^2 \left(\frac{M}{\min_{k,h} C^{k,h}} \right)^3 \frac{1}{\min_{1 \leq m \leq M} q_m} \left(\frac{L}{\mu} \delta_1 \log \left(\frac{1}{\varepsilon} \right) + \frac{M}{\min_{k,h} C^{k,h}} \frac{\delta_2}{\mu^2 \varepsilon} + \frac{M}{\min_{k,h} C^{k,h}} \frac{\sigma^2}{\mu^2 \varepsilon} \right) \right)$ number of devices communications.

2945 THE USE OF LARGE LANGUAGE MODELS (LLMs)

2946 In this work, large language models (LLMs) were used exclusively for spelling edits.