Journal of Machine Learning Research 26 (2025) 1-50 Submitted 1/24; Revised 5/25; Published 6/25

EF21 with Bells & Whistles: Six Algorithmic Extensions of Modern Error Feedback*

Ilyas Fatkhullin ILYAS.FATKHULLIN@AI.ETHZ.CH
Technical University of Munich, Germany
King Abdullah University of Science and Technology, Saudi Arabia
ETH Zurich & ETH AI Center, Switzerland

Igor Sokolov IGOR.SOKOLOV.1@KAUST.EDU.SA
King Abdullah University of Science and Technology, Saudi Arabia

Eduard Gorbunov EDUARD.GORBUNOV@MBZUAI.AC.AE
Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
Moscow Institute of Physics and Technology, Russia

Zhize Li ZHIZELI@SMU.EDU.SG
King Abdullah University of Science and Technology, Saudi Arabia
Singapore Management University, Singapore

Peter Richtárik PETER.RICHTARIK@KAUST.EDU.SA
King Abdullah University of Science and Technology, Saudi Arabia

Editor: Zaid Harchaoui

Abstract

First proposed by Seide et al. (2014) as a heuristic, error feedback (EF) is a very popular mechanism for enforcing convergence of distributed gradient-based optimization methods enhanced with communication compression strategies based on the application of contractive compression operators. However, existing theory of EF relies on very strong assumptions (e.g., bounded gradients), and provides pessimistic convergence rates (e.g., while the best known rate for EF in the smooth nonconvex regime, and when full gradients are compressed, is $O(1/T^{2/3})$, the rate of gradient descent in the same regime is $O(1/T)$). Recently, Richtárik et al. (2021) proposed a new error feedback mechanism, EF21, based on the construction of a Markov compressor induced by a contractive compressor. EF21 removes the aforementioned theoretical deficiencies of EF and at the same time works better in practice. In this work we propose six practical extensions of EF21, all supported by strong convergence theory: partial participation, stochastic approximation, variance reduction, proximal setting, momentum and bidirectional compression. To the best of our knowledge, several of these techniques have not been previously analyzed in combination with EF, and in cases where prior analysis exists—such as for bidirectional compression—our theoretical convergence guarantees significantly improve upon existing results.

Keywords: distributed computing, compressed communication, error feedback.

*. The work was done when E. Gorbunov was a PhD student at MIPT, and I. Fatkhullin was a Master student at TU Munich and a summer intern at KAUST.

1. Introduction

In this paper, we consider the nonconvex distributed optimization problem of the form

$$\min_{x\in\mathbb{R}^d}\left\{f(x)\overset{\text{def}}{=}\frac{1}{n}\sum_{i=1}^{n}f_i(x)\right\}, \tag{1}$$

where n denotes the number of clients/nodes connected with a server/master and client i has an access to the local loss function f_i only. The local loss of each client is allowed to have the online/expectation form

$$f_i(x)=\mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left[f_{\xi_i}(x)\right], \tag{2}$$

or the finite-sum form

$$f_i(x)=\frac{1}{m}\sum_{j=1}^{m}f_{ij}(x). \tag{3}$$

A notable application for problems with such structure is federated learning (Konečný et al., 2016; Kairouz, 2019), where training is performed directly on the clients' devices. In a quest for state-of-the-art performance, machine learning practitioners develop elaborate model architectures and train their models on large data sets. Naturally, in order to make the training at this scale tractable, one needs to rely on distributed computing (Goyal et al., 2017; You et al., 2020). Moreover, massively over-parameterized models have recently shown a remarkable empirical success (Arora et al., 2018). However, the application of these models puts an additional complication on the communication links during training. In order to address this issue, recent research activity and practice focuses on developing distributed optimization methods and systems capitalizing on (deterministic or randomized) *lossy communication compression* techniques to reduce the amount of communication traffic.

A compression mechanism is typically formalized as an operator $\mathcal{C}:\mathbb{R}^d\mapsto\mathbb{R}^d$ mapping hard-to-communicate (e.g., dense) input messages into easy-to-communicate (e.g., sparse) output messages. The operator is allowed to be randomized, and typically operates on models (Khaled and Richtárik, 2019) or on gradients (Alistarh et al., 2017; Beznosikov et al., 2023), both of which can be described as vectors in $\mathbb{R}^d$. Besides sparsification (Alistarh et al., 2018), typical examples of useful compression mechanisms include quantization (Alistarh et al., 2017; Horvóth et al., 2022) and low-rank approximation (Vogels et al., 2019; Safaryan et al., 2022).

There are two large classes of compression operators often studied in the literature: i) *unbiased* compression operators $\mathcal{C}$, meaning that there exists $\omega\geq 0$ such that for all $x\in\mathbb{R}^d$

$$\mathbb{E}\left[\mathcal{C}(x)\right]=x,\quad \mathbb{E}\left[\|\mathcal{C}(x)-x\|^2\right]\quad\leq\quad\omega\|x\|^2; \tag{4}$$

and ii) *biased* compression operators $\mathcal{C}$, meaning that there exists $0<\alpha\leq 1$ such that for all $x\in\mathbb{R}^d$

$$\mathbb{E}\left[\|\mathcal{C}(x)-x\|^2\right]\leq(1-\alpha)\|x\|^2. \tag{5}$$

Note that the latter "biased" class contains the former one, i.e., if $\mathcal{C}$ satisfies (4) with ω, then a scaled version $(1+\omega)^{-1}\mathcal{C}$ satisfies (5) with $\alpha=1/(1+\omega)$. Beyond this inclusion, if used appropriately, biased compressors (such as Top-k sparsifier) often perform better than unbiased ones (such as Rand-k) (Beznosikov et al., 2023). While distributed optimization methods with unbiased compressors (4) are well understood (Alistarh et al., 2017; Khirirat et al., 2018; Mishchenko et al., 2024; Horváth et al., 2019; Li et al., 2020; Li and Richtárik, 2021a; Li and Richtárik, 2020; Islamov et al., 2021; Gorbunov

et al., 2021), *biased* compressors (5) are significantly harder to analyze. One of the main reasons behind this is rooted in the observation that when deployed within distributed gradient descent in a naive way, biased compressors may lead to (even exponential) divergence (Karimireddy et al., 2019; Beznosikov et al., 2023). *Error Feedback* (EF) (or *Error Compensation* (EC))—a technique originally proposed by Seide et al. (2014)—emerged as an empirical fix of this problem. However, this technique remained poorly understood until very recently.

Although several theoretical results were obtained supporting the EF framework in recent years (Stich et al., 2018; Alistarh et al., 2018; Beznosikov et al., 2023; Gorbunov et al., 2020; Qian et al., 2020; Tang et al., 2020; Koloskova et al., 2020; Stich and Karimireddy, 2020), they use strong assumptions (e.g., convexity, bounded gradients, bounded dissimilarity), and do not get $\mathcal{O}(1/\alpha T)$ convergence rates in the smooth nonconvex regime. Very recently, Richtárik et al. (2021) proposed a new EF mechanism called EF21, which uses standard smoothness assumptions only, and also enjoys the desirable $O(1/\alpha T)$ convergence rate for the nonconvex case (in terms of number of communication rounds T this matches the best-known rate $\mathcal{O}((1+\omega/\sqrt{n})/T)$ obtained by Gorbunov et al. (2021) using unbiased compressors), improving the previous $O(1/(\alpha T)^{2/3})$ rate of the standard EF mechanism (Koloskova et al., 2020).

2. Contributions

While Richtárik et al. (2021) propose a new error feedback method, the authors only study their EF21 mechanism in a pure form, without any additional "bells and whistles" which are important in practice. Therefore, it remains elusive whether EF21 method is a standalone technique or it can be enhanced with other related techniques to benefit its potential use in practice. In this paper, we aim to push the EF21 framework beyond its pure form by extending it in several directions of high theoretical and practical importance. In particular, we further enhance the EF21 mechanism with the following six useful and practical algorithmic extensions: *stochastic approximation*, *variance reduction*, *partial participation*, *bidirectional compression*, *momentum*, and *proximal (regularization)*. We do not stop at merely proposing these algorithmic enhancements: we derive *strong convergence results for all of these extensions*. Several of these techniques were never analyzed in conjunction with the original EF mechanism before. This fact reveals the challenges in the analysis of EF-based methods and neccesitates the development of novel analysis techniques. Moreover, in the cases when the mentioned techniques were analyzed with EF-based methods, we obtain new results that are superior in several aspects. See Table 1 for an overview of our results. In summary, our results constitute the new algorithmic and theoretical state-of-the-art in the area of error feedback.

We now briefly comment on each extension proposed in this paper:

◇ **Stochastic approximation.** Vanilla EF21 method requires all clients to compute the exact/full gradient in each round.[1] However, exact gradients are not available in the stochastic/online setting (2), and in the finite-sum setting (3) it is more efficient in practice to use subsampling and work with stochastic gradients instead. In our paper, we extend EF21 to a more general stochastic approximation framework than the simplistic full gradient setting considered in the original paper.

◇ **Variance reduction.** As mentioned above, EF21 relies on full gradient computations at all clients. This incurs a high or unaffordable computational cost, especially when local clients hold large

1. While Richtárik et al. (2021) do consider a stochastic extension of EF21 in their Appendix F, they do not formalize their result, and only consider the simplistic scenario of uniformly bounded variance, which does not in general hold for stochasticity coming from subsampling (Khaled and Richtárik, 2020).

Table 1: Summary of the state-of-the-art complexity results for finding an ε**-stationary point** using error-feedback type methods, where $\varepsilon > 0$ is an accuracy level. That is we aim to find a point $\hat{x}$ such that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \leq \varepsilon^2$, for generally non-convex functions and an ε**-solution**, i.e., such a point $\hat{x}$ that $\mathbb{E}\left[f(\hat{x}) - f(x^*)\right] \leq \varepsilon$, for functions satisfying PŁ-condition. By (computation) complexity we mean the average number of (stochastic) first-order oracle calls needed to find an ε-stationary point ("Compl. (NC)") or ε-solution ("Compl. (PŁ)"). Removing the terms colored in blue from the complexity bounds shown in the table, one can get communication complexity bounds, i.e., the total number of communication rounds needed to find an ε-stationary point ("Compl. (NC)") or ε-solution ("Compl. (PŁ)"). Dependences on the numerical constants, "quality" of the starting point, and smoothness constants are omitted in the complexity bounds. Moreover, dependencies on $\log(1/\varepsilon)$ are also omitted in the column "Compl. (PŁ)". Abbreviations: "BC" = bidirectional compression, "PP" = partial participation; "Mom." = momentum; T = the number of communications rounds needed to find an ε-stationary point; #grads = the number of (stochastic) first-order oracle calls needed to find an ε-stationary point. Notation: α = the compression parameter, α_w and α_M = the compression parameters of worker and master nodes respectively for EF21-BC, $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$ (see Example 1), $\Delta^{\inf} = f^{\inf} - \frac{1}{n}\sum_{i=1}^n \frac{1}{m}\sum_{j=1}^m f_{ij}^{\inf}$ (see Example 2), p = probability of sampling the client in EF21-PP, η = momentum parameter. To the best of our knowledge, combinations of error feedback with partial participation (EF21-PP) and proximal versions of error feedback (EF21-Prox) were never analyzed in the literature.

Setup	Method	Citation	Compl. (NC)	Compl. (PŁ)	Comment
Full grads	EF21	(Richtárik et al., 2021)	$\frac{1}{\alpha\varepsilon^2}$	$\frac{1}{\alpha\mu}$	
Stoch. grads	Choco-SGD	(Koloskova et al., 2020)	$\frac{1}{\varepsilon^2} + \frac{G}{\alpha\varepsilon^3} + \frac{\sigma^2}{n\varepsilon^4}$	N/A	$\|\nabla f_i(x)\| \leq G$
	EF21-SGD	(Richtárik et al., 2021)	$\frac{1}{\alpha\varepsilon^2} + \frac{\sigma^2}{\alpha^3\varepsilon^4}$	$\frac{1}{\alpha\mu} + \frac{\sigma^2}{\mu^2\alpha^3\varepsilon}$	UBV (Ex. 1)
	EF21-SGD	(this work)	$\frac{1}{\alpha\varepsilon^2} + \frac{1+\Delta^{\inf}}{\alpha^3\varepsilon^4}$	$\frac{1}{\alpha\mu} + \frac{1+\Delta^{\inf}}{\mu^2\alpha^3\varepsilon}$	IS (Ex. 2)
	EF21-PAGE	(this work)	$\frac{\sqrt{m}+1/\alpha}{\varepsilon^2} + m$	$\frac{\sqrt{m}+1/\alpha}{\mu} + m$	Finite sum form (3)
PP	EF21-PP	(this work)	$\frac{1}{p\alpha\varepsilon^2}$ (1) $+ \frac{1}{\alpha\varepsilon^2}$	$\frac{1}{p\alpha\mu}$ (1) $+ \frac{1}{\alpha\mu}$	Full grads
BC	DoubleSqueeze	(Tang et al., 2020)	$\frac{1}{\varepsilon^2} + \frac{\Delta}{\varepsilon^3} + \frac{\sigma^2}{n\varepsilon^4}$	N/A	$\mathbb{E}\|\mathcal{C}(x) - x\| \leq \Delta$
	EF21-BC	(this work)	$\frac{1}{\alpha_w\alpha_M\varepsilon^2}$	$\frac{1}{\alpha_w\alpha_M\mu}$	Full grads
Mom.	M-CSER	(Xie et al., 2020)(2)	$\frac{1}{\varepsilon^2} + \frac{G}{(1-\eta)\alpha\varepsilon^3}$	N/A	$\|\nabla f_i(x)\| \leq G$
	EF21-HB	(this work)	$\frac{1}{\varepsilon^2}\left(\frac{1}{1-\eta} + \frac{1}{\alpha}\right)$	N/A	Full grads
Prox	EF21-Prox	(this work)	$\frac{1}{\alpha\varepsilon^2}$	$\frac{1}{\alpha\mu}$ (3)	Full grads

(1) Red term = number of communication rounds, blue term = expected number of gradient computations per client.

(2) Xie et al. (2020) consider Nesterov's momentum. Moreover, they analyzed the version with stochastic gradients, bidirectional compression and local steps. However, the derived result is not better than state-of-the-art ones with either stochastic gradients or bidirectional compression. Therefore, to maintain the table compact, we do not include the results of Xie et al. (2020) in the other parts of the table.

(3) This result is obtained under the generalized PŁ-condition for composite optimization, see Appendix I.2. in (Fatkhullin et al., 2021).

training sets, i.e., if m is very large in (3). One important technique for accelerating convergence is to incorporate a *variance reduction* mechanism, which makes use of stochastic gradient estimates obtained in the previous iterations. To the best of our knowledge, it is an open question whether any EF-type mechanism can be enhanced with variance reduction for non-convex objectives. We answer this question in this work by proposing EF21-PAGE method and developing an analysis based on a new Lyapunov function.

◇ **Partial participation.** Pure EF21 method requires *full participation* of clients for solving problem (1), i.e., in each round, the server needs to communicate with all n clients. However, full participation is usually impractical or very hard to achieve in massively distributed (e.g., federated)

learning problems (Konečný et al., 2016; Cho et al., 2020; Kairouz, 2019; Li and Richtárik, 2021b; Zhao et al., 2021). To remedy this situation, we propose a *partial participation* (PP) variant of EF21, EF21-PP (Algorithm 3), which allows to sample only a random subset of clients at each iteration.

◇ **Bidirectional compression.** In many distributed computing systems the *upstream* of communication of messages is the main bottleneck. However, in other architechtures, the *downstream* communication is also costly (Horvóth et al., 2022; Tang et al., 2020; Philippenko and Dieuleveut, 2020) or even has a fixed bandwidth, which can significantly slow down training. In order to address this issue, we further enhance EF21 method by backward compression and propose EF21-BC (Algorithm 4). Our biderectional compression method carefully employs the Markov compressor based on EF21 on the master and client nodes simultaneously. Moreover, we design a novel analysis for the proposed algorithm, which is reminiscent of the our analysis of EF21-PAGE.

◇ **Momentum.** A very successful and popular technique for enhancing both optimization and generalization is momentum/acceleration (Polyak, 1964; Nesterov, 1983; Lan and Zhou, 2018; Allen-Zhu, 2017; Lan et al., 2019; Li, 2021; Loizou and Richtárik, 2020). Moreover, momentum is a key building block behind the widely-used Adam method (Kingma and Ba, 2014). However, in the context of error feedback, acceleration is notoriously difficult to analyze. For instance, in convex regime, additional full vector communication is needed for the analysis (Qian et al., 2020). In non-convex case, the best-known complexity is acheived by M-CSER method (Xie et al., 2020), which is clearly suboptimal in terms ε, η and α, see Table 1. In this work, we overcome this difficulty by carefully incorporating momentum into EF21. We name the resulting method EF21-HB (Algorithm 5) and offer a simple intuitive proof with improved convergence guarantees.

◇ **Proximal setting.** It is common practice to solve *regularized* versions of empirical risk minimization problems instead of their vanilla variants (Shalev-Shwartz and Ben-David, 2014). Thus we consider the regularized (proximal/composite) problem

$$\min_{x\in\mathbb{R}^d}\left\{\Phi(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n f_i(x) + r(x)\right\}, \tag{6}$$

where $r(x):\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$ is a regularizer, e.g., ℓ_1 regularizer $\|x\|_1$ or ℓ_2 regularizer $\|x\|_2^2$. To broaden the applicability of error feedback to such problems, we propose a proximal variant of EF21 to solve the more general composite problems (6), which leads to our EF21-Prox method (Algorithm 6). Again, we are not aware of any method, which can provably solve problem (6) using the Top-k sparsifier in distributed non-convex setting.

Our theoretical complexity results are summarized in Table 1. We describe each algorithm in detail in Section 4 and present the main results of the convergence analysis for all extensions in Section 5. Proof sketches and formal proofs are deferred to their respective sections in the Appendix. In addition, we also analyze EF21-SGD, EF21-PAGE, EF21-PP, EF21-BC under Polyak-Łojasiewicz (PŁ) condition (Polyak, 1963; Lojasiewicz, 1963) and EF21-Prox under the generalized PŁ-condition (Li and Li, 2018) for composite optimization problems. Due to space limitations, we defer all the details about the analysis under the PŁ-condition to the preliminary version of this work (Fatkhullin et al., 2021) and provide only simplified rates in Table 1. We comment on some preliminary experimental results in Section 6. Additional experiments are presented in Appendix J and in the extended version (Fatkhullin et al., 2025).

3. Notations

We adopt the common conventions $[n] = \{1, \ldots, n\}$ for a set of indicies and $\mathbf{Prob}(\mathcal{A})$ for a probability of event $\mathcal{A}$. Throughout the paper, $\|\cdot\|$ denotes the Euclidean norm $\|\cdot\|_2$ unless otherwise stated. For algorithmic notations, we refer to specific algorithms in the next section. In Appendix A, we summarize all notations used in our theoretical analysis.

4. Methods: Six Algorithmic Extensions

The proposed methods are extensions of EF21, thus they share some features, and are presented in a unified way in Table 3. For all methods, at each iteration, worker i computes the compressed vector c_i^t and sends it to the master. The methods EF21-SGD, EF21-PAGE, EF21-PP differ in how the compressed vectors c_i^t are computed, while the aggregation and parameter update rules are the same:

$$x^{t+1} = x^t - \gamma g^t, \quad g_i^{t+1} = g_i^t + c_i^t,$$
$$g^{t+1} = \frac{1}{n}\sum_{i=1}^n g_i^{t+1} = g^t + \frac{1}{n}\sum_{i=1}^n c_i^t. \tag{7}$$

The methods EF21-BC, EF21-HB, EF21-Prox compute the compressed vectors via $c_i^t = \mathcal{C}(\nabla f_i(x^{t+1}) - g_i^t)$, while the aggregation rule and parameter updates are specific to each method. The pseudocodes of the algorithms are given below and important distinct parts are highlighted in light blue. Below we briefly describe each method.[2]

◇ EF21-SGD**: Error feedback and** SGD**.** EF21-SGD is EF21 with full gradients $\nabla f_i(x^{t+1})$ being replaced by their stochastic estimates $\hat{g}_i(x^{t+1})$ at each node. The pseudocode is given in Algorithm 1. Each client computes $c_i^t = \mathcal{C}(\hat{g}_i(x^{t+1}) - g_i^t)$ and sends this sparsified vector to the server. Despite the simplicity of this extension, it is important for various applications of machine learning and statistics where exact gradients are either unavailable or prohibitively expensive to compute.

Algorithm 1 EF21-SGD

Input: starting point $x^0 \in \mathbb{R}^d$; $g_i^0 \in \mathbb{R}^d$ (known by nodes); $g^0 = \frac{1}{n}\sum_{i=1}^n g_i^0$ (known by master); learning rate $\gamma > 0$
for $t = 0, 1, 2, \ldots, T-1$ **do**
  Master computes $x^{t+1} = x^t - \gamma g^t$ and broadcasts x^{t+1} to all nodes
  for all nodes $i = 1, \ldots, n$ **in parallel do**
    Compute a stochastic gradient $\hat{g}_i(x^{t+1}) = \frac{1}{\tau}\sum_{j=1}^{\tau} \nabla f_{\xi_{ij}^t}(x^{t+1})$
    Compress $c_i^t = \mathcal{C}(\hat{g}_i(x^{t+1}) - g_i^t)$ and send c_i^t to the master
    Update local state $g_i^{t+1} = g_i^t + c_i^t$
  end for
  Master computes $g^{t+1} = \frac{1}{n}\sum_{i=1}^n g_i^{t+1}$ via $g^{t+1} = g^t + \frac{1}{n}\sum_{i=1}^n c_i^t$
end for

2. Note that in this work, we study the effect of each of the 6 proposed extensions separately for pedagogical and clarity reasons. However, in practice, it can be desirable to combine more enhancements with EF21 simultaneously to achieve state-of-the-art performance. In fact, due to the flexibility of our analysis, several of the proposed extensions can be easily combined into one method.

◇ EF21-PAGE**: Error feedback and variance reduction.** In the finite-sum setting (3), it is well known that variance reduced methods have better theoretical guarantees and often perform better than vanilla SGD (Gower et al., 2020). Therefore, we enhance EF21 with variance reduction technique aiming to acheive a stronger combined effect of variance reduction and compressed communication. Specifically, we replace $\nabla f_i(x^{t+1})$ in the formula for c_i^t with the PAGE estimator v_i^{t+1}. With (typically small) probability p this estimator equals the full gradient $v_i^{t+1} = \nabla f_i(x^{t+1})$, and with probability $1-p$ it is set to $v_i^{t+1} = v_i^t + \frac{1}{\tau_i} \sum\limits_{j \in I_i^t} \left(\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)\right)$, where I_i^t is a minibatch of size τ_i. Typically, the number of data points m owned by each client is large, and $p \leq 1/m$ when $\tau_i \equiv 1$. As a result, computation of full gradients rarely happens during the optimization procedure: on average, once in every m iterations only. Although it is possible to use other variance-reduced estimators like in SVRG or SAGA, we use the PAGE-estimator: unlike SVRG or SAGA, PAGE is optimal for smooth nonconvex optimization, and therefore gives the best theoretical guarantees.[3]

Notice that unlike VR-MARINA (Gorbunov et al., 2021), which is a state-of-the-art distributed optimization method designed specifically for unbiased compressors and which *also* uses the PAGE-estimator, by design our EF21-PAGE does not require the communication of full (non-compressed) vectors at all. This is an important property of the algorithm since, in some distributed networks, and especially when d is very large, as is the case in modern over-parameterized deep learning, full vector communication is prohibitive.

Algorithm 2 EF21-PAGE

Input: starting point $x^0 \in \mathbb{R}^d$; $g_i^0, v_i^0 \in \mathbb{R}^d$ for $i = 1, \ldots, n$ (known by nodes); $g^0 = \frac{1}{n}\sum_{i=1}^n g_i^0$ (known by master); learning rate $\gamma > 0$; probabilities $p_i \in (0, 1]$; batch-sizes $1 \leq \tau_i \leq m$
for $t = 0, 1, 2, \ldots, T-1$ **do**
  Master computes $x^{t+1} = x^t - \gamma g^t$
  for all nodes $i = 1, \ldots, n$ **in parallel do**
    Sample $b_i^t \sim \mathrm{Be}(p_i)$
    If $b_i^t = 0$, sample a minibatch of data samples I_i^t with $|I_i^t| = \tau_i$
    $v_i^{t+1} = \begin{cases} \nabla f_i(x^{t+1}) & \text{if } b_i^t = 1, \\ v_i^t + \frac{1}{\tau_i} \sum\limits_{j \in I_i^t} \left(\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)\right) & \text{if } b_i^t = 0 \end{cases}$
    Compress $c_i^t = \mathcal{C}(v_i^{t+1} - g_i^t)$ and send c_i^t to the master
    Update local state $g_i^{t+1} = g_i^t + c_i^t$
  end for
  Master computes $g^{t+1} = \frac{1}{n}\sum_{i=1}^n g_i^{t+1}$ via $g^{t+1} = g^t + \frac{1}{n}\sum_{i=1}^n c_i^t$
end for

◇ EF21-PP**: Error feedback and partial participation.** In this setting, we assume that only a subset of clients is available for computation/communication at each round. We model such situation as follows. First, we select a subset of clients $S_t \subseteq \{1, \ldots, n\}$ randomly such that $\mathbf{Prob}(i \in S_t) = p_i > 0$ for all $i = 1, \ldots, n$, where $\{p_i\}_{i=1}^n$ are unknown probabilities. We allow for an arbitrary sampling strategy of a subset S_t at the master node. The only requirement is that $p_i > 0$ for all

3. We have obtained results for both SVRG and SAGA and indeed, they are worse, and hence we do not include them.

$i = 1, \ldots, n$, which is often referred to as a *proper arbitrary* sampling.[4] Many popular sampling procedures fell into this setting, for instance, independent sampling with/without replacement, τ-nice sampling.[5] After we select a subset S_t, each client from S_t computes $c_i^t = \mathcal{C}(\nabla f_i(x^{t+1}) - g_i^t)$ and communicates this information to the server, while other clients do not participate in the round, which is mathematically equivalent to setting $c_i = 0$. Finally, the server aggregates the communicated vectors and forms a gradient estimator by setting $g_i^{t+1} = g_i^t + c_i^t$ for the clients in S_t and reusing the previous estimate $g_i^{t+1} = g_i^t$ for those nodes which did not take part in this round.

The modified method (Algorithm 3) is called EF21-PP. Note, that all other clients (nodes) $i \notin S_t$ participate neither in the computation nor in communication at iteration t, which can save additional computational effort.

Algorithm 3 EF21-PP (EF21 with partial participation)

Input: starting point $x^0 \in \mathbb{R}^d$; $g_i^0 \in \mathbb{R}^d$ for $i = 1, \ldots, n$ (known by nodes); $g^0 = \frac{1}{n}\sum_{i=1}^n g_i^0$ (known by master); learning rate $\gamma > 0$
for $t = 0, 1, 2, \ldots, T-1$ **do**
  Master computes $x^{t+1} = x^t - \gamma g^t$
  Master samples a subset S_t of nodes ($|S_t| \leq n$) such that $\mathbf{Prob}\,(i \in S_t) = p_i$
  Master broadcasts x^{t+1} to the nodes with $i \in S_t$
  for all nodes $i = 1, \ldots, n$ **in parallel do**
    if $i \in S_t$ **then**
      Compress $c_i^t = \mathcal{C}(\nabla f_i(x^{t+1}) - g_i^t)$ and send c_i^t to the master
      Update local state $g_i^{t+1} = g_i^t + c_i^t$
    end if
    if $i \notin S_t$ **then**
      Do not change local state $g_i^{t+1} = g_i^t$
    end if
  end for
  Master updates $g_i^{t+1} = g_i^t$, $c_i^t = 0$ for $i \notin S_t$
  Master computes $g^{t+1} = \frac{1}{n}\sum_{i=1}^n g_i^{t+1}$ via $g^{t+1} = g^t + \frac{1}{n}\sum_{i=1}^n c_i^t$
end for

$\diamond$ EF21-BC**: Error feedback and bidirectional compression.** We extend EF21 to the case when it is desirable to obtain efficient communication between the clients and the server in *both directions*. We present the formal pseudocode of the method in Algorithm 4. Note that $\mathcal{C}_M$ and $\mathcal{C}_w$ stand for contractive compressors of the type (1) of master and workers respectively. In general, different α_M and α_w are accepted. At each iteration of EF21-BC, clients compute and send to the master node $c_i^t = \mathcal{C}_w(\nabla f_i(x^{t+1}) - \widetilde{g}_i^t)$ and update $\widetilde{g}_i^{t+1} = \widetilde{g}_i^t + c_i^t$ in the usual way, i.e., clients apply EF21 mechanism. The key enhancement in EF21-BC is that the master node in EF21-BC also follows a similar procedure: it computes and broadcasts to clients the compressed vector $b^{t+1} = \mathcal{C}_M(\widetilde{g}^{t+1} - g^t)$ and updates $g^{t+1} = g^t + b^{t+1}$, where $\widetilde{g}^{t+1} = \frac{1}{n}\sum_{i=1}^n \widetilde{g}_i^{t+1}$. Vector g^t is maintained by the master *and*

4. It is natural to focus on *proper* samplings only since otherwise there is a node i, which never communicaties. This would be a critical issue when trying to minimize (1) as we do not assume any similarity between $f_i(\cdot)$.
5. We do not discuss particular sampling strategies here, more details on specific sampling procedures can be found, e.g., in (Qu and Richtárik, 2016).

clients. Therefore, the clients are able to update it via $g^{t+1} = g^t + b^{t+1}$ and compute $x^{t+1} = x^t - \gamma g^t$ once they receive b^{t+1}.

Algorithm 4 EF21-BC (EF21 with bidirectional biased compression)

Input: starting point $x^0 \in \mathbb{R}^d$; g^0, b^0, $\widetilde{g}_i^0 \in \mathbb{R}^d$ for $i = 1, \ldots, n$ (known by nodes); $\widetilde{g}^0 = \frac{1}{n}\sum_{i=1}^n \widetilde{g}_i^0$ (known by master) ; learning rate $\gamma > 0$
for $t = 0, 1, 2, \ldots, T-1$ **do**
  Master updates $x^{t+1} = x^t - \gamma g^t$
  for all nodes $i = 1, \ldots, n$ **in parallel do**
    Update $x^{t+1} = x^t - \gamma g^t$, $g^{t+1} = g^t + b^t$
    compress $c_i^t = \mathcal{C}_w(\nabla f_i(x^{t+1}) - \widetilde{g}_i^t)$, send c_i^t to the master, and
    update local state $\widetilde{g}_i^{t+1} = \widetilde{g}_i^t + c_i^t$
  end for
  Master computes $\widetilde{g}^{t+1} = \frac{1}{n}\sum_{i=1}^n \widetilde{g}_i^{t+1}$ via $\widetilde{g}^{t+1} = \widetilde{g}^t + \frac{1}{n}\sum_{i=1}^n c_i^t$
  compreses $b^{t+1} = \mathcal{C}_M(\widetilde{g}^{t+1} - g^t)$, broadcast b^{t+1} to workers, and
  updates $g^{t+1} = g^t + b^{t+1}$
end for

◇ EF21-HB**: Error feedback with momentum.** We design a momentum (Polyak, 1964) variant of EF21 by computing a moving average estimator based on the vector g^t formed by EF21:

$$\begin{aligned} x^{t+1} &= x^t - \gamma v^t, \quad v^{t+1} = \eta v^t + g^{t+1}, \\ g_i^{t+1} &= g_i^t + c_i^t, \quad g^{t+1} = \frac{1}{n}\sum_{i=1}^n g_i^{t+1} = g^t + \frac{1}{n}\sum_{i=1}^n c_i^t. \end{aligned}$$

The resulting method obtains an improved iteration complexity compared to the current state-of-the-art momentum based method M-CSER in terms of the dependence on ε, η and α, see Table 1. Compared to EF21, its momentum variant EF21-HB has the same complexity (in terms of ε and α), i.e., momentum does not provably improve the convergence rate.[6]

Algorithm 5 EF21-HB

Input: starting point $x^0 \in \mathbb{R}^d$; $g_i^0 \in \mathbb{R}^d$ for $i = 1, \ldots, n$ (known by nodes); $v^0 = g^0 = \frac{1}{n}\sum_{i=1}^n g_i^0$ (known by master); learning rate $\gamma > 0$; momentum parameter $0 \leq \eta < 1$
for $t = 0, 1, 2, \ldots, T-1$ **do**
  Master computes $x^{t+1} = x^t - \gamma v^t$ and broadcasts x^{t+1} to all nodes
  for all nodes $i = 1, \ldots, n$ **in parallel do**
    Compress $c_i^t = \mathcal{C}(\nabla f_i(x^{t+1}) - g_i^t)$ and send c_i^t to the master
    Update local state $g_i^{t+1} = g_i^t + c_i^t$
  end for
  Master computes $g^{t+1} = \frac{1}{n}\sum_{i=1}^n g_i^{t+1}$ via $g^{t+1} = g^t + \frac{1}{n}\sum_{i=1}^n c_i^t$, and $v^{t+1} = \eta v^t + g^{t+1}$
end for

6. Unfortunately, this is a common issue for a wide range of results for momentum methods Loizou and Richtárik (2020). However, it is important to theoretically analyze momentum-extensions such as EF21-HB due to their importance in practice and generalization behaviour.

◇ EF21-Prox**: Error feedback for composite problems.** Finally, we make EF21 applicable to the composite optimization problems (6) by simply taking the prox-operator from the right-hand side of the x^{t+1} update rule (7):

$$x^{t+1} = \operatorname{prox}_{\gamma r}\left(x^t - \gamma g^t\right) \stackrel{\text{def}}{=} \arg\min_{x\in\mathbb{R}^d}\left\{\gamma r(x) + \frac{1}{2}\|x - x^t + \gamma g^t\|^2\right\}.$$

This modification is simple, but, surprisingly, EF21-Prox is the first distributed method with error-feedback that provably converges for composite problems (6). The technical reason for this is that the perturbed iterate analysis of the original EF (Stich et al., 2018; Stich and Karimireddy, 2020) is difficult to extend to the composite/constrainted setting due to additional bias of the proximal operator.

Algorithm 6 EF21-Prox

Input: starting point $x^0 \in \mathbb{R}^d$; $g_i^0 \in \mathbb{R}^d$ for $i = 1, \ldots, n$ (known by nodes); $g^0 = \frac{1}{n}\sum_{i=1}^n g_i^0$ (known by master); learning rate $\gamma > 0$
for $t = 0, 1, 2, \ldots, T-1$ **do**
  Master computes $x^{t+1} = \operatorname{prox}_{\gamma r}\left(x^t - \gamma g^t\right)$
  for all nodes $i = 1, \ldots, n$ **in parallel do**
    Compress $c_i^t = \mathcal{C}(\nabla f_i(x^{t+1}) - g_i^t)$ and send c_i^t to the master
    Update local state $g_i^{t+1} = g_i^t + c_i^t$
  end for
  Master computes $g^{t+1} = \frac{1}{n}\sum_{i=1}^n g_i^{t+1}$ via $g^{t+1} = g^t + \frac{1}{n}\sum_{i=1}^n c_i^t$
end for

5. Theoretical Convergence Results

In this section, we formulate a single corollary derived from the main convergence theorems for our six enhancements of EF21, and formulate the assumptions that we use in the analysis. The complete statements of the theorems and their proofs are provided in the appendices. In Table 1 we compare our new findings with existing results.

5.1 Assumptions

In this subsection, we list and discuss the assumptions that we use in the analysis.

5.1.1 GENERAL ASSUMPTIONS

To derive our convergence results, we invoke the following standard smoothness assumption.

Assumption 1 (Smoothness and lower boundedness) *Every f_i has L_i-Lipschitz gradient, i.e., $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i\|x - y\|$ for all $i \in [n], x, y \in \mathbb{R}^d$, and $f^{\inf} \stackrel{def}{=} \inf_{x\in\mathbb{R}^d} f(x) > -\infty$.*

We also assume that the compression operators used by all algorithms satisfy the following property.

Definition 1 (Contractive compressors) *We say that a (possibly randomized) map $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is a* contractive compression operator*, or simply* contractive compressor*, if there exists a constant $0 < \alpha \leq 1$ such that*

$$\mathbb{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq (1-\alpha)\|x\|^2, \qquad \forall x \in \mathbb{R}^d. \tag{8}$$

We emphasize that we do *not* assume $\mathcal{C}$ to be unbiased. Hence, in particular, our theory works with the popular greedy Top-k sparsifer (Alistarh et al., 2018), which selects the largest k coordinates of the compressed vector in the magnitude.

5.1.2 ADDTIONAL ASSUMPTIONS FOR EF21-SGD

We analyze EF21-SGD under the assumption that local stochastic gradients $\nabla f_{\xi_{ij}^t}(x^t)$ satisfy the following inequality (see Assumption 2 of Khaled and Richtárik (2020)).

Assumption 2 (General assumption for stochastic gradients) *We assume that for all $i = 1, \ldots, n$ and $j \geq 1$, we have $\mathbb{E}\left[\nabla f_{\xi_{ij}^t}(x^t) \mid x^t\right] = \nabla f_i(x^t)$, and there exist parameters $A_i, C_i \geq 0$, $B_i \geq 1$ such that*

$$\mathbb{E}\left[\|\nabla f_{\xi_{ij}^t}(x^t)\|^2 \mid x^t\right] \quad \leq \quad 2A_i\left(f_i(x^t) - f_i^{\inf}\right) + B_i\|\nabla f_i(x^t)\|^2 + C_i, \tag{9}$$

where[7] *$f_i^{\inf} = \inf_{x\in\mathbb{R}^d} f_i(x) > -\infty$.*

Stochastic gradient $\hat{g}_i(x^t)$ is computed using a mini-batch of τ_i independent samples satisfying (9):

$$\hat{g}_i(x^t) \overset{\text{def}}{=} \frac{1}{\tau_i}\sum_{j=1}^{\tau_i} \nabla f_{\xi_{ij}^t}(x^t).$$

Below we provide two examples of stochastic gradients fitting this assumption (for more detail, see (Khaled and Richtárik, 2020)).

Example 1 *Consider $\nabla f_{\xi_{ij}^t}(x^t)$ such that*

$$\mathbb{E}\left[\nabla f_{\xi_{ij}^t}(x^t) \mid x^t\right] = \nabla f_i(x^t) \quad \textit{and} \qquad \mathbb{E}\left[\left\|\nabla f_{\xi_{ij}^t}(x^t) - \nabla f_i(x^t)\right\|^2 \mid x^t\right] \leq \sigma_i^2$$

for some $\sigma_i \geq 0$. Then, due to variance decomposition,(9) holds with $A_i = 0$, $B_i = 0$, $C_i = \sigma_i^2$.

Example 2 *Let $f_i(x) = \frac{1}{m}\sum_{j=1}^m f_{ij}(x)$, f_{ij} be L_{ij}-smooth and $f_{ij}^{\inf} = \inf_{x\in\mathbb{R}^d} f_{ij}(x) > -\infty$. Following Gower et al. (2019), we consider a stochastic reformulation*

$$f_i(x) \quad = \quad \mathbb{E}_{v_i\sim\mathcal{D}_i}\left[f_{v_i}(x)\right] = \mathbb{E}_{v_i\sim\mathcal{D}_i}\left[\frac{1}{m}\sum_{j=1}^m f_{v_{ij}}(x)\right],$$

where $\mathbb{E}_{v_i\sim\mathcal{D}_i}\left[v_{ij}\right] = 1$. One can show (see Proposition 2 of Khaled and Richtárik (2020)) that under the assumption that $\mathbb{E}_{v_i\sim\mathcal{D}_i}\left[v_{ij}^2\right]$ is finite for all j stochastic gradient $\nabla f_{\xi_{ij}^t}(x^t) = \nabla f_{v_i^t}(x^t)$ with v_i^t sampled from $\mathcal{D}_i$ satisfies (9) with $A_i = \max_j L_{ij}\mathbb{E}_{v_i\sim\mathcal{D}_i}\left[v_{ij}^2\right]$, $B_i = 1$, $C_i = 2A_i\Delta_i^{\inf}$, where $\Delta_i^{\inf} = \frac{1}{m}\sum_{j=1}^m (f_i^{\inf} - f_{ij}^{\inf})$. In particular, if $\mathbf{Prob}(\nabla f_{\xi_{ij}^t}(x^t) = \nabla f_{ij}(x^t)) = \frac{L_{ij}}{\sum_{l=1}^m L_{il}}$, then $A_i = \overline{L}_i = \frac{1}{m}\sum_{j=1}^m L_{ij}$, $B_i = 1$, and $C_i = 2A_i\Delta_i^{\inf}$.

7. When $A_i = 0$ one can ignore the first term in the right-hand side of (9), i.e., assumption $\inf_{x\in\mathbb{R}^d} f_i(x) > -\infty$ is not required in this case.

5.1.3 ADDITIONAL ASSUMPTIONS FOR EF21-PAGE

In the analysis of EF21-PAGE, we rely on the following assumption.

Assumption 3 (Average $\mathcal{L}$-smoothness) *Let every f_i have the form (3). Assume that for all $t \geq 0$, $i = 1, \ldots, n$, and batch I_i^t (of size τ_i), the minibatch stochastic gradients difference $\widetilde{\Delta}_i^t \overset{def}{=} \frac{1}{\tau_i} \sum_{j \in I_i^t} (\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t))$ computed on the node i, satisfies $\mathbb{E}\left[\widetilde{\Delta}_i^t \mid x^t, x^{t+1}\right] = \Delta_i^t$ and*

$$\mathbb{E}\left[\left\|\widetilde{\Delta}_i^t - \Delta_i^t\right\|^2 \mid x^t, x^{t+1}\right] \leq \frac{\mathcal{L}_i^2}{\tau_i} \|x^{t+1} - x^t\|^2 \tag{10}$$

with some $\mathcal{L}_i \geq 0$, where $\Delta_i^t \overset{def}{=} \nabla f_i(x^{t+1}) - \nabla f_i(x^t)$. We also define $\widetilde{\mathcal{L}}^2 \overset{def}{=} \frac{1}{n} \sum_{i=1}^n \frac{(1-p_i)\mathcal{L}_i^2}{\tau_i}$.

This assumption is satisfied for many standard/popular sampling strategies. For example, if I_i^t is a full batch, then $\mathcal{L}_i = 0$. Another example is *uniform sampling* on $\{1, \ldots, m\}$, and each f_{ij} is L_{ij}-smooth. In this regime, one may verify that $\mathcal{L}_i \leq \max_{1 \leq j \leq m} L_{ij}$.

5.1.4 ADDITIONAL ASSUMPTIONS FOR GLOBAL CONVERGENCE

We now introduce an additional assumption, which enables us to obtain (global) convergence results for the function value.

Assumption 4 (Polyak-Łojasiewicz) *There exists $\mu > 0$ such that $f(x) - f(x^\star) \leq \frac{1}{2\mu} \|\nabla f(x)\|^2$ for all $x \in \mathbb{R}^d$, where $x^\star \in \arg\min_{x \in \mathbb{R}^d} f(x) \neq \emptyset$.*

The results under this assumption (and its generalization to composite case) are briefly summarized in Table 1 in the column "Compl. (PŁ)". The detailed statements of the results are deferred to the preliminary version of this work (Fatkhullin et al., 2021).

5.2 Main results

Below, we formulate the corollary establishing the complexities for each method. The complete version of this result is formulated and rigorously derived for each method in the appendix. We also include the proof sketch for each result in the corresponding sections.

Corollary 2 *Suppose that Assumption 1 holds. Then, there exist appropriate choices of parameters for EF21-PP, EF21-BC, EF21-HB, EF21-Prox such that the number of communication rounds T and the (expected) number of gradient computations at each node $\#grad$ for these methods to find an ε-stationary point, i.e., a point $\hat{x}^T$ such that $\mathbb{E}[\|\nabla f(\hat{x}^T)\|^2] \leq \varepsilon^2$ for EF21-PP, EF21-BC, EF21-HB and $\mathbb{E}[\|\mathcal{G}_\gamma(\hat{x}^T)\|^2] \leq \varepsilon^2$ for EF21-Prox, where $\mathcal{G}_\gamma(x) = 1/\gamma \left(x - \mathrm{prox}_{\gamma r}(x - \gamma \nabla f(x))\right)$, are*

$$\begin{aligned}
\text{EF21-PP:} \quad & T = \mathcal{O}\left(\frac{\widetilde{L}\delta^0}{p\alpha\varepsilon^2}\right), \quad \#grad = \mathcal{O}\left(\frac{\widetilde{L}\delta^0}{\alpha\varepsilon^2}\right) \\
\text{EF21-BC:} \quad & T = \#grad = \mathcal{O}\left(\frac{\widetilde{L}\delta^0}{\alpha_w \alpha_M \varepsilon^2}\right) \\
\text{EF21-HB:} \quad & T = \#grad = \mathcal{O}\left(\frac{\widetilde{L}\delta^0}{\varepsilon^2}\left(\frac{1}{\alpha} + \frac{1}{1-\eta}\right)\right) \\
\text{EF21-Prox:} \quad & T = \#grad = \mathcal{O}\left(\frac{\widetilde{L}\delta^0}{\alpha\varepsilon^2}\right),
\end{aligned}$$

where $\widetilde{L} \overset{def}{=} \sqrt{\frac{1}{n}\sum_{i=1}^n L_i^2}$, $\delta_0 \overset{def}{=} f(x^0) - f^{\inf}$ *(for* EF21-Prox $\delta^0 = \Phi(x^0) - \Phi^{inf}$*), p is the probability of sampling the client in* EF21-PP*, α_w and α_M are contraction factors for compressors applied on the workers' and the master's sides respectively in* EF21-BC*, and $\eta \in [0, 1)$ is the momentum parameter in* EF21-HB*.*

If Assumptions 1 and 2 in the setup from Example 1 hold, then there exist appropriate choices of parameters for EF21-SGD *such that the corresponding T and the averaged number of gradient computations at each node $\overline{\#grad}$ are*

$$\text{EF21-SGD:} \qquad T = \mathcal{O}\left(\frac{\widetilde{L}\delta^0}{\alpha\varepsilon^2}\right), \qquad \overline{\#grad} = \mathcal{O}\left(\frac{\widetilde{L}\delta^0}{\alpha\varepsilon^2} + \frac{\widetilde{L}\delta^0\sigma^2}{\alpha^3\varepsilon^4}\right),$$

where $\sigma = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$.

If Assumptions 1 and 3 hold, then there exist appropriate choices of parameters for EF21-PAGE *such that the corresponding T and $\overline{\#grad}$ are*

$$\text{EF21-PAGE:} \qquad T = \mathcal{O}\left(\frac{(\widetilde{L}+\widetilde{\mathcal{L}})\delta^0}{\alpha\varepsilon^2} + \frac{\sqrt{m}\widetilde{\mathcal{L}}\delta^0}{\varepsilon^2}\right), \qquad \overline{\#grad} = \mathcal{O}\left(m + \frac{(\widetilde{L}+\widetilde{\mathcal{L}})\delta^0}{\alpha\varepsilon^2} + \frac{\sqrt{m}\widetilde{\mathcal{L}}\delta^0}{\varepsilon^2}\right),$$

where $\widetilde{\mathcal{L}} = \sqrt{\frac{1-p}{n}\sum_{i=1}^n \mathcal{L}_i^2}$, $\tau_i \equiv \tau = 1$.

Discussion of results and comparison to prior work:

• For EF21-PP and EF21-Prox, none of previous error feedback methods work on these two settings (partial participation and proximal/composite case). Thus, we provide the *first* convergence results for them. Moreover, we show that the gradient (computation) complexity for both EF21-PP and EF21-Prox is $\mathcal{O}(1/\alpha\varepsilon)$, matching the original vanilla EF21. It means that we extend EF21 to both settings for free.

• For EF21-BC, we show $\mathcal{O}(1/\alpha_w\alpha_M\varepsilon^2)$ complexity result, which naturally extends $\mathcal{O}(1/\alpha_w\varepsilon^2)$ complexity to the case when compression is also applied by server. The most related method, which applies a biased compression in both directions, is DoubleSqueeze of Tang et al. (2020). This algorithm acheives only $\mathcal{O}(\Delta/\varepsilon^3)$, moreover, it uses a strong assumption on the compressors ($\mathbb{E}\left[\|\mathcal{C}(x) - x\|\right] \leq \Delta$), which is not satisfied for practically interesing examples such as Top-K. Therefore, our analysis improves upon the previous result by acheving better rates and using a more flexible class of compressors.

• Our iteration complexity for EF21-HB is of order $\mathcal{O}(1/\varepsilon^2)$. In contrast, the previous result of M-CSER is $\mathcal{O}(G/\varepsilon^3)$ and its analysis requires an additional bounded gradient assumption. Moreover, we improve the dependence on momentum and contraction parameters by splitting the product of $(1-\eta)^{-1}$ and α^{-1} into the sum.

• For EF21-SGD and EF21-PAGE, we want to reduce the gradient complexity by using (variance-reduced) stochastic gradients instead of full gradient in the vanilla EF21. Note that σ^2 and $\Delta^{\inf}$ in EF21-SGD could be much smaller than G in Choco-SGD, while σ^2 and $\Delta^{\inf}$ are often dimension-free parameters (particularly, they are very small if the functions/data samples are similar). Thus, for high dimensional problems (e.g., deep neural networks), EF21-SGD can be better than Choco-SGD. Besides, in the finite-sum case (3), especially if the number of data samples m on each client is not very large, then EF21-PAGE is much better since its sample complexity is $\mathcal{O}(\sqrt{m}/\varepsilon^2)$ while for EF21-SGD it is of order $\mathcal{O}(\sigma^2/\varepsilon^4)$. We can also observe that the sample complexities of EF21-SGD and EF21-PAGE do not have the linear speedup in the number of nodes (i.e., the devision by n) as it is present in

(distributed) SGD (Khaled and Richtárik, 2020). We experimentally verify the tightness of our rates w.r.t. the number of clients n; these results are presented in Appendix J of the extended version of this work (Fatkhullin et al., 2025).

5.3 Proof sketches

In this section we provide insights into convergence analysis for several of our extensions: EF21-PP, EF21-PAGE and EF21-HB. Proof sketches for EF21-SGD, EF21-BC and EF21-Prox are deferred to Appendix and can be found in the corresponding sections.

Partial participation. The idea of our analysis of EF21-PP is to develop a recursion on the error term $G_i^{t+1} \overset{\text{def}}{=} \left\|g_i^{t+1} - \nabla f_i(x^{t+1})\right\|^2$ analogous to that in original EF21 analysis (see Lemma 3). We do this by conditioning on the events $\mathcal{A}_i$ that node i is sampled to participate in communication round, i.e., $i \in S_t$. That is, we consider the following two terms:

$$\mathbb{E}\left[G_i^{t+1} \mid i \in S_t\right] \qquad \text{and} \quad \mathbb{E}\left[G_i^{t+1} \mid i \notin S_t\right].$$

The strategy for controlling these two terms is different. In the first case, when node i participates in training, progress is made toward improving the accuracy of the g_i^{t+1} estimator. In the second case, an additional cost arises because node i skips the communication round. If we can bound each term above efficiently, we can continue by computing the full (unconditional) expectation using conditional expectations derived from events $\mathcal{A}_i$ and its complement.

$$\mathbb{E}\left[G_i^{t+1}\right] = p_i\,\mathbb{E}\left[G_i^{t+1} \mid i \in S_t\right] + (1-p_i)\,\mathbb{E}\left[G_i^{t+1} \mid i \notin S_t\right],$$

Finally, combining the established recursion on the expected error term, $\mathbb{E}\left[G_i^{t+1}\right]$, with the standard descent lemma and performing a careful calculation of the final communication and iteration complexities allows us to establish convergence guarantees. The full proof is deferred to Appendix E.

Variance reduction. The key strategy for analyzing EF21-PAGE involves splitting the error into two parts,

$$\left\|\nabla f_i(x^t) - g_i^t\right\|^2 \le 2\left\|\nabla f_i(x^t) - v_i^t\right\|^2 + 2\left\|v_i^t - g_i^t\right\|^2,$$

and bounding each term separately. The first term corresponds to an error due to variance reduction, and the second term is related to the EF21 mechanism with the compressor. The strategy of controlling the first term (see Lemma 11) is similar to the analysis of error deviation of PAGE estimator in (Li et al., 2021). However, controlling the second term (see Lemma 12) is more involved due to the interplay between the two errors. Indeed, while both sequences v_i^t and g_i^t change dynamically, the key challenge is to efficiently control the accumulated error from both and build up a recursion of type

$$\mathbb{E}\left[\left\|v_i^{t+1} - g_i^{t+1}\right\|^2\right] \le (1-\theta)\mathbb{E}\left[\left\|v_i^t - g_i^t\right\|^2\right] + C_1\mathbb{E}\left[\left\|\nabla f_i(x^t) - v_i^t\right\|^2\right] + C_2\mathbb{E}\left[\left\|x^{t+1} - x^t\right\|^2\right],$$

where $\theta \in (0,1)$ is a contraction factor, and $C_1, C_2 > 0$ are constants determined by problem structure and algorithm's parameters. Once this recursion is established, it is combined with a similar recursion for $\mathbb{E}\left[\left\|\nabla f_i(x^t) - v_i^t\right\|^2\right]$ and descent lemma (see Lemma 21), which results in the following Lyapunov function

$$f(x^t) - f^{\inf} + \frac{\gamma}{\theta n}\sum_{i=1}^n \left\|\nabla f_i(x^t) - v_i^t\right\|^2 + \frac{C_3}{n}\sum_{i=1}^n \left\|v_i^t - g_i^t\right\|^2,$$

where $\gamma > 0$ is step-size and $C_3 > 0$. See Appendix D for more details.

Heavy ball momentum. The key idea of the convergence analysis of EF21-HB is in line with (Yang et al., 2016; Liu et al., 2020), where an additional virtual sequence $\left\{z^t\right\}_{t\geq 0}$ is defined as

$$z^{t+1} = x^{t+1} - \frac{\gamma\eta}{1-\eta}v^t,$$

where $\gamma > 0$ is step-size and $\eta \in [0, 1)$ is momentum parameter of Algorithm 5. The main challenge is to control the error term introduced by the EF21 mechanism with a contractive compressor, while accounting for the momentum step, which replaces the simple gradient descent step used in the original EF21. The error term due to compression is controlled as in the EF21 analysis by showing

$$\mathbb{E}\left[\left\|\nabla f_i(x^{t+1}) - g_i^{t+1}\right\|^2\right] \leq (1-\theta)\mathbb{E}\left[\left\|\nabla f_i(x^t) - g_i^t\right\|^2\right] + \beta L_i^2 \mathbb{E}\left[\left\|x^{t+1} - x^t\right\|^2\right], \tag{11}$$

where $\theta \in (0, 1)$ represents a contraction factor, and $\beta > 0$ depends on contraction factor α. However, the Lyapunov function used in EF21-HB differs from that of EF21:

$$f(z^t) - f^{\inf} + \frac{C}{n}\sum_{i=1}^n \left\|g_i^t - \nabla f_i(x^t)\right\|^2,$$

where $C > 0$, since a virtual sequence z^t appears in function value in the first term instead of x^t. This difference causes a technical difficulty in controlling the last term in (11) since it is different from $\left\|z^{t+1} - z^t\right\|^2$ involved in the descent type lemma for EF21-HB. We overcome this challenge by relating these two terms after summation as

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|x^{t+1} - x^t\right\|^2\right] \leq 2(1+4\eta^2)\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|z^{t+1} - z^t\right\|^2\right].$$

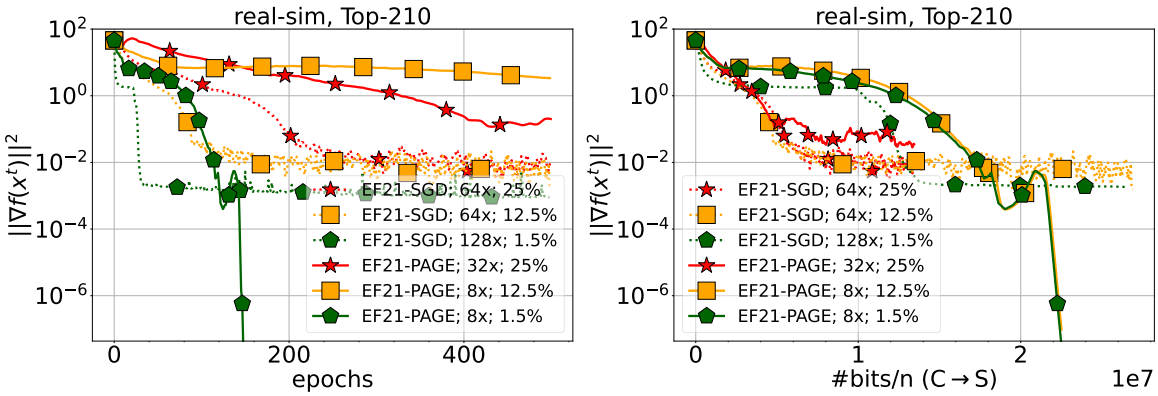
We refer to Lemma 25 in Appendix G of the extended version of this paper for a rigorous proof (Fatkhullin et al., 2025).

6. Experiments

In this section, we consider a logistic regression problem with a non-convex regularizer, i.e., $f(x) = \frac{1}{N}\sum\limits_{i=1}^N \log\left(1 + \exp\left(-b_i a_i^\top x\right)\right) + \lambda\sum\limits_{j=1}^d \frac{x_j^2}{1+x_j^2}$, where $a_i \in \mathbb{R}^d, b_i \in \{-1, 1\}$ are the training data, and $\lambda > 0$ is the regularization parameter, which is set to $\lambda = 0.1$ in all experiments. We use $n = 20$ for experiments $1, 3$ and $n = 100$ for experiment 2, and split datapoints heterogeneously. In all algorithms involving compression, we use Top-k (Alistarh et al., 2017) as a canonical example of contractive compressor $\mathcal{C}$, and fix the compression ratio $k/d \approx 0.01$, where d is the number of features in the data set. For all algorithms, at each iteration we compute the squared norm of the exact/full gradient for comparison of the methods performance. We terminate our algorithms either if they reach the certain number of iterations or the following stopping criterion is satisfied: $\left\|\nabla f(x^t)\right\|^2 \leq 10^{-7}$. We tune the step-sizes for each method individually and report the best one based on the minimal number of bits required to acheive the desired accuracy. We refer the reader to Appendix J for more detailed experimental setup, and additional experiments, including other proposed methods such as

EF21-HB and EF21-BC.[8] The main goal of the following numerical experiments is to illustrate our key theoretical findings. This way we further motivate the proposed algorithmic enhancements of EF21.

Experiment 1: Fast convergence with variance reduction. In our first experiment, we showcase the computation and communication benefit of EF21-PAGE (Alg. 2) over EF21-SGD. Figure 1 illustrates that, in all cases, EF21-PAGE perfectly reduces the accumulated variance and converges to the desired tolerance, whereas EF21-SGD is stuck at some accuracy level. Moreover, EF21-PAGE turns out to be surprisingly efficient with small batchsizes (eg, 1.5% of the local data) both in terms of the number of epochs and the # bits sent to the server per client. Interestingly, for most data sets, a further increase of batchsize does not considerably improve the convergence.



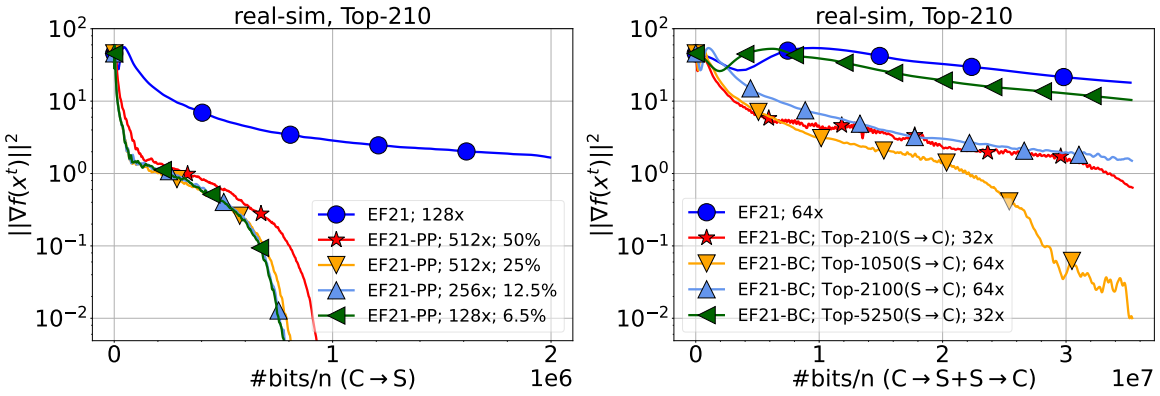
(a) Convergence in epochs.

(b) Convergence in terms of total number of bits sent from **C**lients to the **S**erver divided by n.

Figure 1: Comparison of EF21-PAGE and EF21-SGD with tuned parameters. By $1\times, 2\times, 4\times$ (and so on) we indicate that the stepsize was set to a multiple of the largest stepsize predicted by theory for EF21. By 25%, 12.5% and 1.5% we refer to batchsizes equal $\lfloor 0.25N_i \rfloor$, $\lfloor 0.125N_i \rfloor$ and $\lfloor 0.015N_i \rfloor$ for all clients $i = 1, \ldots, n$, where N_i denotes the size of local data set.

Experiment 2: On the effect of partial participation of clients. This experiment shows that EF21-PP (Alg. 3) has potential to reduce communication cost. For this comparison, we consider $n = 100$, and apply a different data partitioning, see Table 6 in Appendix J for more details. It is predicted by our theory (Corollary 2) that, in terms of the number of iterations/communication rounds, *partial participation* slows down the convergence of EF21 by a fraction of participating clients. However, since for EF21-PP the communications are considerably cheaper it is able to outperform EF21 in terms of the number of bits sent to the server per client on average (see Figure 2a).

8. Implementation of all our algorithms is publicly available at https://github.com/IgorSokoloff/ef21_bw_experiements_source_code.



(a) Convergence in terms of total number of bits sent from **C**lients to the **S**erver divided by n.

(b) Convergence measured by (bits sent from **C**lients to the **S**erver + bits from **S**erver to **C**lients) $/n$.

Figure 2: Comparison of EF21, EF21-PP and EF21-BC with tuned parameters. By $1\times, 2\times, 4\times$ (and so on) we indicate that the stepsize was set to a multiple of the largest stepsize predicted by theory for EF21.

Experiment 3: On the advantages of bidirectional biased compression. Our next experiment demonstrates that the application of the **S**erver $\to$ **C**lients compression in EF21-BC (Alg. 4) improves convergence in terms of bits to be transmitted from **S**erver $\to$ **C**lients and **C**lients $\to$ **S**erver together. Indeed, Figure 2b illustrates that EF21-BC outperforms EF21 in terms of the total number of bits communicated even when communicating only $5\% - 15\%$ of data.[9] Note that EF21 communicates full vectors from the **S**erver $\to$ **C**lients, which slows down communication at each round.

We refer to the full version of this paper for more ablation studies with different data sets and deep learning experiments (Fatkhullin et al., 2021).

Conclusion

This work extends the capabilities of EF21 by introducing six practical enhancements: partial participation, stochastic approximation, variance reduction, proximal settings, momentum, and bidirectional compression. These extensions address key limitations of earlier error feedback methods, offering improved theoretical guarantees and practical performance. While our results highlight significant progress, further work is needed to explore their full potential in real-world scenarios, such as federated learning in highly heterogeneous regimes. We hope these contributions will inspire continued advancements in communication-efficient optimization.

Acknowledgments

The authors would like to thank the anonymous reviewers and the handling editor for their constructive feedback and suggestions, which helped improve the quality and clarity of this paper. This work was supported by funding from King Abdullah University of Science and Technology (KAUST) Baseline

9. The range $5\% - 15\%$ comes from the fractions k/d for each data set and this observation is consistent accross several data sets.

Research Scheme. I. Fatkhullin is partially funded by ETH AI Center Doctoral Fellowship. The work of E. Gorbunov was partially supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Moscow Institute of Physics and Technology dated November 1, 2021 No. 70-2021-00138.

Table of Contents

Table 2: Summary of frequently used notations in the proofs.

Algorithm	Notation
for all algorithms	$G_i^t = \left\|g_i^t - \nabla f_i(x^t)\right\|^2$, $G^t = \frac{1}{n}\sum_{i=1}^n G_i^t$
EF21 EF21-SGD EF21-PP	$R^t = \left\|x^{t+1} - x^t\right\|^2$, $\delta^t = f(x^t) - f^{\inf}$
EF21-PAGE	$R^t = \left\|x^{t+1} - x^t\right\|^2$, $\delta^t = f(x^t) - f^{\inf}$, $P_i^t = \left\|\nabla f_i(x^t) - v_i^t\right\|^2$, $V_i^t = \left\|v_i^t - g_i^t\right\|^2$, $P^t = \frac{1}{n}\sum_{i=1}^n P_i^t$, $V^t = \frac{1}{n}\sum_{i=1}^n V_i^t$
EF21-BC	$R^t = \left\|x^{t+1} - x^t\right\|^2$, $\delta^t = f(x^t) - f^{\inf}$, $P_i^t = \left\|\widetilde{g}_i^t - \nabla f_i(x^t)\right\|^2$, $P^t = \frac{1}{n}\sum_{i=1}^n P_i^t$
EF21-HB	$R^t = (1-\eta)^2\left\|z^{t+1} - z^t\right\|^2$, $\delta^t = f(z^t) - f^{\inf}$
EF21-Prox	$R^t = \left\|x^{t+1} - x^t\right\|^2$, $\Phi(x) = f(x) + r(x)$, $\delta^t = \Phi(x^t) - \Phi^{inf}$, $\mathcal{G}_\gamma(x) = \frac{1}{\gamma}\left(x - \mathrm{prox}_{\gamma r}(x - \gamma\nabla f(x))\right)$

Appendix A. Tables with Notations and Methods

Table 2 summarizes the most frequently used notations in our analysis. Additionally, we comment on the main quantities here. Following Richtárik et al. (2021), we denote the deviation of EF21 estimator g_i^t from the local gradient by $G_i^t \overset{\text{def}}{=} \left\|\nabla f_i(x^t) - g_i^t\right\|^2$, and by $G^t \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n G_i^t$, the average of this quantity over multiple nodes. This notation is common for analysis of all algorithms in this work. In Section D for EF21-PAGE, it is useful to split the deviation G_i^t further and define the corresponding deviations of variance reduced estimator v_i^t from exact local gradient $\nabla f_i(x^t)$, $P_i^t \overset{\text{def}}{=} \left\|\nabla f_i(x^t) - v_i^t\right\|^2$, and v_i^t from EF21 estimator g_i^t, $V_i^t \overset{\text{def}}{=} \left\|v_i^t - g_i^t\right\|^2$. Similarly, in Section F for EF21-BC, it is helpful to consider additionally the deviation between EF21 estimator on the clients and the exact gradient $P_i^t \overset{\text{def}}{=} \left\|\widetilde{g}_i^t - \nabla f_i(x^t)\right\|^2$ and the deviation between EF21 estimator on the server and the average of EF21 estimators on the clients $V^t \overset{\text{def}}{=} \left\|\frac{1}{n}\sum_{i=1}^n \widetilde{g}_i^t - g^t\right\|^2$.

For analysis of most algorithms, we define $\delta^t \overset{\text{def}}{=} f(x^t) - f^{\inf}$,[10] $R^t \overset{\text{def}}{=} \left\|x^{t+1} - x^t\right\|^2$. In the analysis of EF21-HB, it is useful to modify this notation to $\delta^t \overset{\text{def}}{=} f(z^t) - f^{\inf}$ and $R^t \overset{\text{def}}{=} (1-\eta)^2\left\|z^{t+1} - z^t\right\|^2$, where $\{z^t\}_{t\geq 0}$ is the sequence of virtual iterates introduced in Section G.

10. If, additionally, Assumption 4 holds, then $f^{\inf}$ can be replaced by $f(x^\star)$ for $x^\star \in \arg\min_{x\in\mathbb{R}^d} f(x) \neq \emptyset$.

Table 3: Description of the methods developed and analyzed in the paper. For the ease of comparison, we also provide a description of EF21. In all methods only compressed vectors c_i^t are transmitted from workers to the master and the master broadcasts non-compressed iterates x^{t+1} (except EF21-BC, where the master broadcasts compressed vector b^{t+1}). Initialization of g_i^0, $i = 1, \ldots, n$ can be arbitrary (possibly randomized). One possible choice is $g_i^0 = \mathcal{C}(\nabla f_i(x^0))$. The pseudocodes for each method are given in the appendix.

Method	EF21-	c_i^t	Comment
$x^{t+1} = x^t - \gamma g^t$, $g^t = \frac{1}{n}\sum_{i=1}^n g_i^t$, $g_i^{t+1} = g_i^t + c_i^t$	n/a Alg. 7	$\mathcal{C}(\nabla f_i(x^{t+1}) - g_i^t)$	
	SGD Alg. 1	$\mathcal{C}(\hat{g}_i(x^{t+1}) - g_i^t)$	$\hat{g}_i(x^{t+1})$ satisfies As. 2
	PAGE Alg. 2	$\mathcal{C}(v_i^{t+1} - g_i^t)$	$b_i^t \sim \text{Be}(p)$, $v_i^{t+1} = \nabla f_i(x^{t+1})$, if $b_i^t = 1$, $v_i^{t+1} = v_i^t + \frac{1}{\tau_i}\sum_{j\in I_i^t} \nabla f_{ij}(x^{t+1}) - \frac{1}{\tau_i}\sum_{j\in I_i^t} \nabla f_{ij}(x^t)$, if $b_i^t = 0$, I_i^t is a minibatch, $\lvert I_i^t\rvert = \tau_i$
	PP Alg. 3	$\mathcal{C}(\nabla f_i(x^{t+1}) - g_i^t)$ 0	if $i \in S_t$ if $i \notin S_t$
$x^{t+1} = x^t - \gamma g^t$, $g^{t+1} = g^t + b^{t+1}$, $b^{t+1} = \mathcal{C}_M(\widetilde{g}^{t+1} - g^t)$, $\widetilde{g}^{t+1} = \frac{1}{n}\sum_{i=1}^n \widetilde{g}_i^{t+1}$, $\widetilde{g}_i^{t+1} = \widetilde{g}_i^t + c_i^t$	BC Alg. 4	$\mathcal{C}_w(\nabla f_i(x^{t+1}) - \widetilde{g}_i^t)$	Master broadcasts b^{t+1}; $\mathcal{C}_w$ used by workers, $\mathcal{C}_M$ used by master
$x^{t+1} = x^t - \gamma v^t$, $v^{t+1} = \eta v^t + g^{t+1}$, $g^{t+1} = \frac{1}{n}\sum_{i=1}^n g_i^{t+1}$, $g_i^{t+1} = g_i^t + c_i^t$	HB Alg. 5	$\mathcal{C}(\nabla f_i(x^{t+1}) - g_i^t)$	$\eta \in [0, 1)$ is momentum parameter
$x^{t+1} = \text{prox}_{\gamma r}\left(x^t - \gamma g^t\right)$, $g^{t+1} = \frac{1}{n}\sum_{i=1}^n g_i^{t+1}$, $g_i^{t+1} = g_i^t + c_i^t$	Prox Alg. 6	$\mathcal{C}(\nabla f_i(x^{t+1}) - g_i^t)$	For problem (6); $\text{prox}_{\gamma r}(x)$ is defined in (39)

Appendix B. EF21

For completeness, we provide here the pseudocode and the detailed convergence proof for EF21 (Richtárik et al., 2021).

Algorithm 7 EF21

Input: starting point $x^0 \in \mathbb{R}^d$; $g_i^0 \in \mathbb{R}^d$ for $i = 1, \ldots, n$ (known by nodes); $g^0 = \frac{1}{n}\sum_{i=1}^n g_i^0$ (known by master); learning rate $\gamma > 0$
for $t = 0, 1, 2, \ldots, T-1$ **do**
  Master computes $x^{t+1} = x^t - \gamma g^t$ and broadcasts x^{t+1} to all nodes
  for all nodes $i = 1, \ldots, n$ **in parallel do**
    Compress $c_i^t = \mathcal{C}(\nabla f_i(x^{t+1}) - g_i^t)$ and send c_i^t to the master
    Update local state $g_i^{t+1} = g_i^t + c_i^t$
  end for
  Master computes $g^{t+1} = \frac{1}{n}\sum_{i=1}^n g_i^{t+1}$ via $g^{t+1} = g^t + \frac{1}{n}\sum_{i=1}^n c_i^t$
end for

Lemma 3 *Let $\mathcal{C}$ be a contractive compressor, then for all $i = 1, \ldots, n$*

$$\mathbb{E}\left[G_i^{t+1}\right] \leq (1-\theta)\mathbb{E}\left[G_i^t\right] + \beta L_i^2 \mathbb{E}\left[\left\|x^{t+1} - x^t\right\|^2\right], \text{ and} \tag{12}$$

$$\mathbb{E}\left[G^{t+1}\right] \leq (1-\theta)\mathbb{E}\left[G^t\right] + \beta \widetilde{L}^2 \mathbb{E}\left[\left\|x^{t+1} - x^t\right\|^2\right], \tag{13}$$

where $\theta \overset{def}{=} 1-(1-\alpha)(1+s)$, $\quad \beta \overset{def}{=} (1-\alpha)\left(1+s^{-1}\right)$ for any $s > 0$.

Proof Define $W^t \overset{\text{def}}{=} \{g_1^t, \ldots, g_n^t, x^t, x^{t+1}\}$, then

$$\begin{aligned}
\mathbb{E}\left[G_i^{t+1}\right] &= \mathbb{E}\left[\mathbb{E}\left[G_i^{t+1} \mid W^t\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\|g_i^{t+1} - \nabla f_i(x^{t+1})\right\|^2 \mid W^t\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\|g_i^t + \mathcal{C}(\nabla f_i(x^{t+1}) - g_i^t) - \nabla f_i(x^{t+1})\right\|^2 \mid W^t\right]\right] \\
&\overset{(8)}{\leq} (1-\alpha)\mathbb{E}\left[\left\|\nabla f_i(x^{t+1}) - g_i^t\right\|^2\right] \\
&\overset{(i)}{\leq} (1-\alpha)(1+s)\mathbb{E}\left[\left\|\nabla f_i(x^t) - g_i^t\right\|^2\right] \\
&\qquad +(1-\alpha)\left(1+s^{-1}\right)\mathbb{E}\left[\left\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right\|^2\right] \qquad (14) \\
&\overset{(ii)}{\leq} (1-\alpha)(1+s)\mathbb{E}\left[\left\|\nabla f_i(x^t) - g_i^t\right\|^2\right] \\
&\qquad +(1-\alpha)\left(1+s^{-1}\right)L_i^2\mathbb{E}\left[\left\|x^{t+1} - x^t\right\|^2\right] \\
&\overset{(iii)}{\leq} (1-\theta)\mathbb{E}\left[\left\|\nabla f_i(x^t) - g_i^t\right\|^2\right] + \beta L_i^2 \mathbb{E}\left[\left\|x^{t+1} - x^t\right\|^2\right],
\end{aligned}$$

where (i) follows by Young's inequality (41), (ii) holds by Assumption 1, and in (iii) we apply the definition of θ and β. Averaging the above inequalities over $i = 1, \ldots, n$, we obtain (13). ■

Theorem 4 *Let Assumption 1 hold, and let the stepsize in Algorithm 7 be set as*

$$0 < \gamma \leq \left(L + \widetilde{L}\sqrt{\frac{\beta}{\theta}} \right)^{-1}. \tag{15}$$

Fix $T \geq 1$ and let $\hat{x}^T$ be chosen from the iterates $x^0, x^1, \ldots, x^{T-1}$ uniformly at random. Then

$$\mathbb{E}\left[\left\| \nabla f(\hat{x}^T) \right\|^2 \right] \leq \frac{2\left(f(x^0) - f^{inf}\right)}{\gamma T} + \frac{\mathbb{E}\left[G^0\right]}{\theta T}, \tag{16}$$

where $\widetilde{L} = \sqrt{\frac{1}{n}\sum_{i=1}^n L_i^2}$, $\theta = 1 - (1-\alpha)(1+s)$, $\beta = (1-\alpha)\left(1 + s^{-1}\right)$ for any $s > 0$.

Proof According to our notation, for Algorithm 7 $R^t = \left\| x^{t+1} - x^t \right\|^2$. By Lemma 3, we have

$$\mathbb{E}\left[G^{t+1}\right] \quad \leq \quad (1-\theta)\,\mathbb{E}\left[G^t\right] + \beta\widetilde{L}^2\mathbb{E}\left[R^t\right]. \tag{17}$$

Next, using Lemma 21 and Jensen's inequality (42), we obtain the bound

$$\begin{aligned} f(x^{t+1}) &\leq f(x^t) - \frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right) R^t + \frac{\gamma}{2}\left\| \frac{1}{n}\sum_{i=1}^n \left(g_i^t - \nabla f_i(x^t)\right) \right\|^2 \\ &\leq f(x^t) - \frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right) R^t + \frac{\gamma}{2}\frac{1}{n}\sum_{i=1}^n \left\| g_i^t - \nabla f_i(x^t) \right\|^2 \\ &= f(x^t) - \frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right) R^t + \frac{\gamma}{2} G^t. \end{aligned} \tag{18}$$

Subtracting $f^{\inf}$ from both sides of the above inequality, taking expectation and using the notation $\delta^t = f(x^t) - f^{\inf}$, we get

$$\mathbb{E}\left[\delta^{t+1}\right] \quad \leq \quad \mathbb{E}\left[\delta^t\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\mathbb{E}\left[R^t\right] + \frac{\gamma}{2}\mathbb{E}\left[G^t\right]. \tag{19}$$

Then by adding (19) with a $\frac{\gamma}{2\theta}$ multiple of (17) we obtain

$$\begin{aligned} \mathbb{E}\left[\delta^{t+1}\right] + \frac{\gamma}{2\theta}\mathbb{E}\left[G^{t+1}\right] &\leq \mathbb{E}\left[\delta^t\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\mathbb{E}\left[R^t\right] + \frac{\gamma}{2}\mathbb{E}\left[G^t\right] \\ &\quad + \frac{\gamma}{2\theta}\left(\beta\widetilde{L}^2\mathbb{E}\left[R^t\right] + (1-\theta)\mathbb{E}\left[G^t\right]\right) \\ &= \mathbb{E}\left[\delta^t\right] + \frac{\gamma}{2\theta}\mathbb{E}\left[G^t\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \\ &\quad - \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{\gamma}{2\theta}\beta\widetilde{L}^2\right)\mathbb{E}\left[R^t\right] \\ &\leq \mathbb{E}\left[\delta^t\right] + \frac{\gamma}{2\theta}\mathbb{E}\left[G^t\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right]. \end{aligned}$$

The last inequality follows from the bound $\gamma^2 \frac{\beta \widetilde{L}^2}{\theta} + L\gamma \leq 1$, which holds because of Lemma 20 and our assumption on the stepsize. By summing up inequalities for $t = 0, \ldots, T-1$, and rearranging we get (16), since $\hat{x}^T$ is chosen from $x^0, x^1, \ldots, x^{T-1}$ uniformly at random. ■

Corollary 5 *Let assumptions of Theorem 4 hold,*

$$\begin{aligned} g_i^0 &= \nabla f_i(x^0), \qquad i = 1, \ldots, n, \\ \gamma &= \left(L + \widetilde{L}\sqrt{\beta/\theta}\right)^{-1}. \end{aligned}$$

Then, after T iterations/communication rounds of EF21 *we have $\mathbb{E}\left[\left\|\nabla f(\hat{x}^T)\right\|^2\right] \leq \varepsilon^2$. It requires*

$$T = \#grad = \mathcal{O}\left(\frac{\widetilde{L}\delta^0}{\alpha\varepsilon^2}\right)$$

iterations/communications rounds/gradint computations at each node, where $\widetilde{L} = \sqrt{\frac{1}{n}\sum_{i=1}^n L_i^2}$, $\delta^0 = f(x^0) - f^{inf}$.

Proof Since $g_i^0 = \nabla f_i(x^0)$, $i = 1, \ldots, n$, we have $G^0 = 0$ and by Theorem 4

$$\begin{aligned} \#\text{grad} &= T \overset{(i)}{\leq} \frac{2\delta^0}{\gamma\varepsilon^2} \overset{(ii)}{\leq} \frac{2\delta^0}{\varepsilon^2}\left(L + \widetilde{L}\sqrt{\frac{\beta}{\theta}}\right) \overset{(iii)}{\leq} \frac{2\delta^0}{\varepsilon^2}\left(L + \widetilde{L}\left(\frac{2}{\alpha} - 1\right)\right) \\ &\leq \frac{2\delta^0}{\varepsilon^2}\left(L + \frac{2\widetilde{L}}{\alpha}\right) \overset{(iv)}{\leq} \frac{2\delta^0}{\varepsilon^2}\left(\frac{\widetilde{L}}{\alpha} + \frac{2\widetilde{L}}{\alpha}\right) = \frac{6\widetilde{L}\delta^0}{\alpha\varepsilon^2}, \end{aligned}$$

where in (i) is due to the rate (16) given by Theorem 4. In (ii) we plug in the stepsize, in (iii) we use Lemma 22, and (iv) follows by the inequalities $\alpha \leq 1$, and $L \leq \widetilde{L}$. ■

Appendix C. Stochastic Gradients

In this section, we study the extension of EF21 to the case when stochastic gradients are used instead of full gradients. The main idea of the proof is to design an analogous recursion as in Lemma 3 for the EF21 error term

$$G_i^{t+1} = \left\|g_i^{t+1} - \nabla f_i(x^{t+1})\right\|^2,$$

where

$$g_i^{t+1} = g_i^t + \mathcal{C}(\hat{g}_i(x^{t+1}) - g_i^t), \qquad \hat{g}_i(x^{t+1}) = \frac{1}{\tau}\sum_{j=1}^{\tau} \nabla f_{\xi_{ij}^t}(x^{t+1}).$$

However, due to additional noise from sampling stochastic gradients, extra error terms occur. The goal of the next lemma is to efficiently control such error terms uing Young's inequality several times and applying Assumption 2 on stochastic gradients.

As in the previous section, we use notations $G^t = \frac{1}{n}\sum_{i=1}^n G_i^t$, $G_i^t = \left\|\nabla f_i(x^t) - g_i^t\right\|^2$.

Lemma 6 *Let Assumptions 1 and 2 hold. Then for all $t \geq 0$ and all constants $\rho, \nu > 0$ EF21-SGD satisfies*

$$\begin{aligned}\mathbb{E}\left[G^{t+1}\right] &\leq (1-\hat{\theta})\mathbb{E}\left[G^t\right] + \hat{\beta}_1\widetilde{L}^2\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right] \\ &\quad + \widetilde{A}\hat{\beta}_2\mathbb{E}\left[f(x^{t+1}) - f^{\inf}\right] + \widetilde{C}\hat{\beta}_2, \end{aligned} \tag{20}$$

where $\hat{\theta} \overset{def}{=} 1-(1-\alpha)(1+\rho)(1+\nu)$, $\hat{\beta}_1 \overset{def}{=} 2(1-\alpha)(1+\rho)\left(1+\frac{1}{\nu}\right)$, $\hat{\beta}_2 \overset{def}{=} 2(1-\alpha)(1+\rho)\left(1+\frac{1}{\nu}\right)+\left(1+\frac{1}{\rho}\right)$, $\widetilde{A} = \max_{i=1,\ldots,n}\frac{2(A_i+L_i(B_i-1))}{\tau_i}$, $\widetilde{C} = \frac{1}{n}\sum\limits_{i=1}^n\left(\frac{2(A_i+L_i(B_i-1))}{\tau_i}\left(f^{\inf}-f_i^{\inf}\right)+\frac{C_i}{\tau_i}\right)$.

Proof Applying Young's inequality with parameter $\rho > 0$

$$\begin{aligned}\mathbb{E}\left[G_i^{t+1}\right] &= \mathbb{E}\left[\left\|g_i^{t+1} - \nabla f_i(x^{t+1})\right\|^2\right] \leq (1+\rho)\mathbb{E}\left[\left\|\mathcal{C}\left(\hat{g}_i(x^{t+1}) - g_i^t\right) - \left(\hat{g}_i(x^{t+1}) - g_i^t\right)\right\|^2\right] \\ &\quad + \left(1+\frac{1}{\rho}\right)\mathbb{E}\left[\left\|\hat{g}_i(x^{t+1}) - \nabla f_i(x^{t+1})\right\|^2\right] \\ &\leq (1-\alpha)(1+\rho)\mathbb{E}\left[\left\|g_i^t - \hat{g}_i(x^{t+1})\right\|^2\right] + \left(1+\frac{1}{\rho}\right)\mathbb{E}\left[\left\|\hat{g}_i(x^{t+1}) - \nabla f_i(x^{t+1})\right\|^2\right].\end{aligned}$$

Further applying Young's inequality with parameters $\nu > 0$ and $s = 2$

$$\begin{aligned}\mathbb{E}\left[G_i^{t+1}\right] &\leq (1-\alpha)(1+\rho)(1+\nu)\mathbb{E}\left[\left\|g_i^t - \nabla f_i(x^t)\right\|^2\right] \\ &\quad + 2(1-\alpha)(1+\rho)\left(1+\frac{1}{\nu}\right)\mathbb{E}\left[\left\|\nabla f_i(x^{t+1}) - \hat{g}_i(x^{t+1})\right\|^2\right] \\ &\quad + 2(1-\alpha)(1+\rho)\left(1+\frac{1}{\nu}\right)\mathbb{E}\left[\left\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right\|^2\right] \\ &\quad + \left(1+\frac{1}{\rho}\right)\mathbb{E}\left[\left\|\hat{g}_i(x^{t+1}) - \nabla f_i(x^{t+1})\right\|^2\right] \\ &\leq (1-\hat{\theta})\mathbb{E}\left[G_i^t\right] + \hat{\beta}_1 L_i^2\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right] + \hat{\beta}_2\mathbb{E}\left[\left\|\hat{g}_i(x^{t+1}) - \nabla f_i(x^{t+1})\right\|^2\right],\end{aligned}$$

where we introduced $\hat{\theta} \overset{\text{def}}{=} 1-(1-\alpha)(1+\rho)(1+\nu)$, $\hat{\beta}_1 \overset{\text{def}}{=} 2(1-\alpha)(1+\rho)\left(1+\frac{1}{\nu}\right)$, $\hat{\beta}_2 \overset{\text{def}}{=} 2(1-\alpha)(1+\rho)\left(1+\frac{1}{\nu}\right)+\left(1+\frac{1}{\rho}\right)$. Next we use independence of $\nabla f_{\xi_{ij}^t}(x^t)$, variance decomposition, and (9) to estimate the last term:

$$
\begin{aligned}
\mathbb{E}\left[G_i^{t+1}\right] &\leq (1-\hat{\theta})\mathbb{E}\left[G_i^t\right]+\hat{\beta}_1 L_i^2\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right] \\
&\quad +\frac{\hat{\beta}_2}{\tau_i^2}\sum_{j=1}^{\tau_i}\mathbb{E}\left[\left\|\nabla f_{\xi_{ij}^t}(x^{t+1})-\nabla f_i(x^{t+1})\right\|^2\right] \\
&= (1-\hat{\theta})\mathbb{E}\left[G_i^t\right]+\hat{\beta}_1 L_i^2\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right] \\
&\quad +\frac{\hat{\beta}_2}{\tau_i^2}\sum_{j=1}^{\tau_i}\left(\mathbb{E}\left[\left\|\nabla f_{\xi_{ij}^t}(x^{t+1})\right\|^2\right]-\mathbb{E}\left[\left\|\nabla f_i(x^{t+1})\right\|^2\right]\right) \\
&\overset{(9)}{\leq} (1-\hat{\theta})\mathbb{E}\left[G_i^t\right]+\hat{\beta}_1 L_i^2\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right] \\
&\quad +\frac{2A_i\hat{\beta}_2}{\tau_i}\mathbb{E}\left[f_i(x^{t+1})-f_i^{\inf}\right]+\frac{\hat{\beta}_2(B_i-1)}{\tau_i}\mathbb{E}\left[\left\|\nabla f_i(x^{t+1})\right\|^2\right]+\frac{C_i\hat{\beta}_2}{\tau_i} \\
&\leq (1-\hat{\theta})\mathbb{E}\left[G_i^t\right]+\hat{\beta}_1 L_i^2\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right] \\
&\quad +\frac{2(A_i+L_i(B_i-1))\hat{\beta}_2}{\tau_i}\mathbb{E}\left[f_i(x^{t+1})-f_i^{\inf}\right]+\frac{C_i\hat{\beta}_2}{\tau_i}.
\end{aligned}
$$

Averaging the obtained inequality for $i=1,\ldots,n$ we get

$$
\begin{aligned}
\mathbb{E}\left[G^{t+1}\right] &\leq (1-\hat{\theta})\mathbb{E}\left[G^t\right]+\hat{\beta}_1\widetilde{L}^2\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right] \\
&\quad +\frac{1}{n}\sum_{i=1}^n\left(\frac{2(A_i+L_i(B_i-1))\hat{\beta}_2}{\tau_i}\mathbb{E}\left[f_i(x^{t+1})-f_i^{\inf}\right]+\frac{C_i\hat{\beta}_2}{\tau_i}\right) \\
&\leq (1-\hat{\theta})\mathbb{E}\left[G^t\right]+\hat{\beta}_1\widetilde{L}^2\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right] \\
&\quad +\frac{1}{n}\sum_{i=1}^n\left(\frac{2(A_i+L_i(B_i-1))\hat{\beta}_2}{\tau_i}\mathbb{E}\left[f_i(x^{t+1})-f^{\inf}\right]\right) \\
&\quad +\frac{\hat{\beta}_2}{n}\sum_{i=1}^n\left(\frac{2(A_i+L_i(B_i-1))}{\tau_i}\left(f^{\inf}-f_i^{\inf}\right)+\frac{C_i}{\tau_i}\right) \\
&\leq (1-\hat{\theta})\mathbb{E}\left[G^t\right]+\hat{\beta}_1\widetilde{L}^2\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right]+\widetilde{A}\hat{\beta}_2\mathbb{E}\left[f(x^{t+1})-f^{\inf}\right]+\widetilde{C}\hat{\beta}_2.
\end{aligned}
$$

■

Theorem 7 *Let Assumptions 1 and 2 hold, and let the stepsize in Algorithm 1 be set as*

$$
0<\gamma\leq\left(L+\widetilde{L}\sqrt{\frac{\hat{\beta}_1}{\hat{\theta}}}\right)^{-1}, \tag{21}
$$

where $\widetilde{L} = \sqrt{\frac{1}{n}\sum_{i=1}^n L_i^2}$, $\hat{\theta} \overset{def}{=} 1-(1-\alpha)(1+\rho)(1+\nu)$, $\hat{\beta}_1 \overset{def}{=} 2(1-\alpha)(1+\rho)\left(1+\frac{1}{\nu}\right)$, *and* $\rho, \nu > 0$ *are some positive numbers. Assume that batchsizes* $\tau_1, \ldots, \tau_i$ *are such that* $\frac{\gamma\widetilde{A}\hat{\beta}_2}{2\hat{\theta}} < 1$, *where* $\widetilde{A} = \max_{i=1,\ldots,n} \frac{2(A_i+L_i(B_i-1))}{\tau_i}$ *and* $\hat{\beta}_2 \overset{def}{=} 2(1-\alpha)(1+\rho)\left(1+\frac{1}{\nu}\right) + \left(1+\frac{1}{\rho}\right)$. *Fix* $T \geq 1$ *and let* $\hat{x}^T$ *be chosen from the iterates* $x^0, x^1, \ldots, x^{T-1}$ *with following probabilities:*

$$\mathbf{Prob}\left\{\hat{x}^T = x^t\right\} = \frac{w_t}{W_T}, \quad w_t = \left(1 - \frac{\gamma\widetilde{A}\hat{\beta}_2}{2\hat{\theta}}\right)^t, \quad W_T = \sum_{t=0}^T w_t.$$

Then

$$\mathbb{E}\left[\left\|\nabla f(\hat{x}^T)\right\|^2\right] \leq \frac{2(f(x^0)-f^{\inf})}{\gamma T\left(1-\frac{\gamma\widetilde{A}\hat{\beta}_2}{2\hat{\theta}}\right)^T} + \frac{\mathbb{E}\left[G^0\right]}{\hat{\theta}T\left(1-\frac{\gamma\widetilde{A}\hat{\beta}_2}{2\hat{\theta}}\right)^T} + \frac{\widetilde{C}\hat{\beta}_2}{\hat{\theta}}, \tag{22}$$

where $\widetilde{C} = \frac{1}{n}\sum\limits_{i=1}^n \left(\frac{2(A_i+L_i(B_i-1))}{\tau_i}\left(f^{\inf}-f_i^{\inf}\right) + \frac{C_i}{\tau_i}\right)$.

Proof We notice that inequality (19) holds for EF21-SGD as well, i.e., we have

$$\mathbb{E}\left[\delta^{t+1}\right] \leq \mathbb{E}\left[\delta^t\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] - \left(\frac{1}{2\gamma}-\frac{L}{2}\right)\mathbb{E}\left[R^t\right] + \frac{\gamma}{2}\mathbb{E}\left[G^t\right].$$

Summing up the above inequality with a $\frac{\gamma}{2\hat{\theta}}$ multiple of (20), we derive

$$\begin{aligned}
\mathbb{E}\left[\delta^{t+1} + \frac{\gamma}{2\hat{\theta}}G^{t+1}\right] &\leq \mathbb{E}\left[\delta^t\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] - \left(\frac{1}{2\gamma}-\frac{L}{2}\right)\mathbb{E}\left[R^t\right] + \frac{\gamma}{2}\mathbb{E}\left[G^t\right] \\
&\quad + \frac{\gamma}{2\hat{\theta}}(1-\hat{\theta})\mathbb{E}\left[G^t\right] + \frac{\gamma}{2\hat{\theta}}\hat{\beta}_1\widetilde{L}^2\mathbb{E}\left[R^t\right] \\
&\quad + \frac{\gamma}{2\hat{\theta}}\widetilde{A}\hat{\beta}_2\mathbb{E}\left[\delta^{t+1}\right] + \frac{\gamma}{2\hat{\theta}}\widetilde{C}\hat{\beta}_2 \\
&\leq \frac{\gamma\widetilde{A}\hat{\beta}_2}{2\hat{\theta}}\mathbb{E}\left[\delta^{t+1}\right] + \mathbb{E}\left[\delta^t + \frac{\gamma}{2\hat{\theta}}G^t\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + \frac{\gamma}{2\hat{\theta}}\widetilde{C}\hat{\beta}_2 \\
&\quad - \left(\frac{1}{2\gamma}-\frac{L}{2}-\frac{\gamma\hat{\beta}_1\widetilde{L}^2}{2\hat{\theta}}\right)\mathbb{E}\left[R^t\right] \\
&\overset{(21)}{\leq} \frac{\gamma\widetilde{A}\hat{\beta}_2}{2\hat{\theta}}\mathbb{E}\left[\delta^{t+1}\right] + \mathbb{E}\left[\delta^t + \frac{\gamma}{2\hat{\theta}}G^t\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + \frac{\gamma}{2\hat{\theta}}\widetilde{C}\hat{\beta}_2,
\end{aligned}$$

where $\hat{\theta} \overset{\text{def}}{=} 1-(1-\alpha)(1+\rho)(1+\nu)$, $\hat{\beta}_1 \overset{\text{def}}{=} 2(1-\alpha)(1+\rho)\left(1+\frac{1}{\nu}\right)$, $\hat{\beta}_2 \overset{\text{def}}{=} 2(1-\alpha)(1+\rho)\left(1+\frac{1}{\nu}\right) + \left(1+\frac{1}{\rho}\right)$, and $\rho, \nu > 0$ are some positive numbers. Next, we rearrange the terms

$$\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \leq \frac{2}{\gamma}\left(\mathbb{E}\left[\delta^t + \frac{\gamma}{2\hat{\theta}}G^t\right] - \left(1-\frac{\gamma\widetilde{A}\hat{\beta}_2}{2\hat{\theta}}\right)\mathbb{E}\left[\delta^{t+1} + \frac{\gamma}{2\hat{\theta}}\mathbb{E}\left[G^{t+1}\right]\right]\right) + \frac{\widetilde{C}\hat{\beta}_2}{\hat{\theta}},$$

sum up the obtained inequalities for $t = 0, 1, \ldots, T$ with weights w_t/W_T, and use the definition of $\hat{x}^T$

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla f(\hat{x}^T)\right\|^2\right] &= \frac{1}{W_K}\sum_{t=0}^{T} w_t \mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \\ &\leq \frac{2}{\gamma W_T}\sum_{t=0}^{T}\left(w_t\mathbb{E}\left[\delta^t + \frac{\gamma}{2\hat{\theta}}G^t\right] - w_{t+1}\mathbb{E}\left[\delta^{t+1} + \frac{\gamma}{2\hat{\theta}}\mathbb{E}\left[G^{t+1}\right]\right]\right) + \frac{\widetilde{C}\hat{\beta}_2}{\hat{\theta}} \\ &\leq \frac{2\delta^0}{\gamma W_T} + \frac{\mathbb{E}\left[G^0\right]}{\hat{\theta}W_T} + \frac{\widetilde{C}\hat{\beta}_2}{\hat{\theta}}.\end{aligned}$$

Finally, we notice $W_T = \sum\limits_{t=0}^{T} w_t \geq (T+1)\min\limits_{t=0,1,\ldots,T} w_t > T\left(1 - \frac{\gamma\widetilde{A}\hat{\beta}_2}{2\hat{\theta}}\right)^T$ that finishes the proof. ■

Corollary 8 *Let assumptions of Theorem 7 hold, $\rho = \alpha/2$, $\nu = \alpha/4$,*

$$\begin{aligned}\gamma &= \frac{1}{L + \widetilde{L}\sqrt{\frac{\hat{\beta}_1}{\hat{\theta}}}}, \\ \tau_i &= \left\lceil \max\left\{1, \frac{2T\gamma\left(A_i + L_i(B_i-1)\right)\hat{\beta}_2}{\hat{\theta}}, \frac{8\left(A_i + L_i(B_i-1)\right)\hat{\beta}_2}{\hat{\theta}\varepsilon^2}\delta_i^{\inf}, \frac{4C_i\hat{\beta}_2}{\hat{\theta}\varepsilon^2}\right\}\right\rceil, \\ T &= \left\lceil \max\left\{\frac{16\delta^0}{\gamma\varepsilon^2}, \frac{8\mathbb{E}\left[G^0\right]}{\hat{\theta}\varepsilon^2}\right\}\right\rceil,\end{aligned}$$

where $\delta_i^{\inf} = f^{\inf} - f_i^{\inf}$, $\delta^0 = f(x^0) - f^{\inf}$. Then, after T iterations of EF21-SGD *we have $\mathbb{E}\left[\left\|\nabla f(\hat{x}^T)\right\|^2\right] \leq \varepsilon^2$. It requires*

$$T = \mathcal{O}\left(\frac{\widetilde{L}\delta^0 + \mathbb{E}\left[G^0\right]}{\alpha\varepsilon^2}\right)$$

iterations/communications rounds,

$$\begin{aligned}\#grad_i &= \tau_i T = \mathcal{O}\Bigg(\frac{\widetilde{L}\delta^0 + \mathbb{E}\left[G^0\right]}{\alpha\varepsilon^2} + \frac{\left(\widetilde{L}\delta^0 + \mathbb{E}\left[G^0\right]\right)\left(\hat{A}_i(\delta^0 + \delta_i^{\inf}) + C_i\right)}{\alpha^3\varepsilon^4} \\ &\qquad + \frac{(\widetilde{L}\delta^0 + \mathbb{E}\left[G^0\right])\hat{A}_i\mathbb{E}\left[G^0\right]}{\alpha^2(\alpha L + \widetilde{L})\varepsilon^4}\Bigg)\end{aligned}$$

stochastic oracle calls for worker i, and

$$\begin{aligned}\overline{\#grad} &= \frac{1}{n}\sum_{i=1}^{n}\tau_i T \\ &= \mathcal{O}\Bigg(\frac{\widetilde{L}\delta^0 + \mathbb{E}\left[G^0\right]}{\alpha\varepsilon^2} + \frac{1}{n}\sum_{i=1}^{n}\frac{\left(\widetilde{L}\delta^0 + \mathbb{E}\left[G^0\right]\right)\left(\hat{A}_i(\delta^0 + \delta_i^{\inf}) + C_i\right)}{\alpha^3\varepsilon^4} \\ &\qquad + \frac{1}{n}\sum_{i=1}^{n}\frac{(\widetilde{L}\delta^0 + \mathbb{E}\left[G^0\right])\hat{A}_i\mathbb{E}\left[G^0\right]}{\alpha^2(\alpha L + \widetilde{L})\varepsilon^4}\Bigg)\end{aligned}$$

stochastic oracle calls per worker on average, where $\hat{A}_i = A_i + L_i(B_i - 1)$.

Proof The given choice of τ_i ensures that $\left(1 - \frac{\gamma\widetilde{A}\hat{\beta}_2}{2\hat{\theta}}\right)^T = \mathcal{O}(1)$ and $\widetilde{C}\hat{\beta}_2/\hat{\theta} \leq \varepsilon/2$. Next, the choice of T ensures that the right-hand side of (22) is smaller than ε. Finally, after simple computation we get the expression for $\tau_i T$. ■

Corollary 9 *Consider the setting described in Example 1. Let assumptions of Theorem 7 hold,* $\rho = \alpha/2$, $\nu = \alpha/4$,

$$\gamma = \frac{1}{L + \widetilde{L}\sqrt{\frac{\hat{\beta}_1}{\hat{\theta}}}}, \quad \tau_i = \left\lceil \max\left\{1, \frac{4\sigma_i^2\hat{\beta}_2}{\hat{\theta}\varepsilon^2}\right\}\right\rceil, \quad T = \left\lceil \max\left\{\frac{16\delta^0}{\gamma\varepsilon^2}, \frac{8\mathbb{E}\left[G^0\right]}{\hat{\theta}\varepsilon^2}\right\}\right\rceil,$$

where $\delta^0 = f(x^0) - f^{\inf}$. *Then, after* T *iterations of* EF21-SGD *we have* $\mathbb{E}\left[\left\|\nabla f(\hat{x}^T)\right\|^2\right] \leq \varepsilon^2$. *It requires*

$$T = \mathcal{O}\left(\frac{\widetilde{L}\delta^0 + \mathbb{E}\left[G^0\right]}{\alpha\varepsilon^2}\right)$$

iterations/communications rounds,

$$\#grad_i \quad = \quad \tau_i T = \mathcal{O}\left(\frac{\widetilde{L}\delta^0 + \mathbb{E}\left[G^0\right]}{\alpha\varepsilon^2} + \frac{\left(\widetilde{L}\delta^0 + \mathbb{E}\left[G^0\right]\right)\sigma_i^2}{\alpha^3\varepsilon^4}\right)$$

stochastic oracle calls for worker i, *and*

$$\overline{\#grad} \quad = \quad \frac{1}{n}\sum_{i=1}^n \tau_i T = \mathcal{O}\left(\frac{\widetilde{L}\delta^0 + \mathbb{E}\left[G^0\right]}{\alpha\varepsilon^2} + \frac{\left(\widetilde{L}\delta^0 + \mathbb{E}\left[G^0\right]\right)\sigma^2}{\alpha^3\varepsilon^4}\right)$$

stochastic oracle calls per worker on average, where $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$.

Corollary 10 *Consider the setting described in Example 2. Let assumptions of Theorem 7 hold,* $\rho = \alpha/2$, $\nu = \alpha/4$,

$$\begin{aligned}
\gamma &= \frac{1}{L + \widetilde{L}\sqrt{\frac{\hat{\beta}_1}{\hat{\theta}}}}, \\
\tau_i &= \left\lceil \max\left\{1, \frac{2T\gamma\overline{L}_i\hat{\beta}_2}{\hat{\theta}}, \frac{8\overline{L}_i\hat{\beta}_2}{\hat{\theta}\varepsilon^2}\delta_i^{\inf}, \frac{8\overline{L}_i\Delta_i^{\inf}\hat{\beta}_2}{\hat{\theta}\varepsilon^2}\right\}\right\rceil, \\
T &= \left\lceil \max\left\{\frac{16\delta^0}{\gamma\varepsilon^2}, \frac{8\mathbb{E}\left[G^0\right]}{\hat{\theta}\varepsilon^2}\right\}\right\rceil,
\end{aligned}$$

where $\delta_i^{\inf} = f^{\inf} - f_i^{\inf}$, $\delta^0 = f(x^0) - f^{\inf}$, $\overline{L}_i = \frac{1}{m}\sum_{j=1}^m L_{ij}$, $\Delta_i^{\inf} = \frac{1}{m}\sum_{j=1}^m (f_i^{\inf} - f_{ij}^{\inf})$. *Then, after* T *iterations of* EF21-SGD *we have* $\mathbb{E}\left[\left\|\nabla f(\hat{x}^T)\right\|^2\right] \leq \varepsilon^2$. *It requires*

$$T = \mathcal{O}\left(\frac{\widetilde{L}\delta^0 + \mathbb{E}\left[G^0\right]}{\alpha\varepsilon^2}\right)$$

iterations/communications rounds,

$$\#grad_i \quad = \quad \tau_i T = \mathcal{O}\Bigg(\frac{\widetilde{L}\delta^0 + \mathbb{E}\left[G^0\right]}{\alpha\varepsilon^2} + \frac{\left(\widetilde{L}\delta^0 + \mathbb{E}\left[G^0\right]\right)\left(\overline{L}_i(\delta^0 + \delta_i^{\inf}) + \overline{L}_i\Delta_i^{\inf}\right)}{\alpha^3\varepsilon^4} + \frac{(\widetilde{L}\delta^0 + \mathbb{E}\left[G^0\right])\overline{L}_i\mathbb{E}\left[G^0\right]}{\alpha^2(\alpha L + \widetilde{L})\varepsilon^4}\Bigg)$$

stochastic oracle calls for worker i, and

$$\begin{aligned}\overline{\#grad} \quad &= \quad \frac{1}{n}\sum_{i=1}^n \tau_i T \\ &= \quad \mathcal{O}\Bigg(\frac{\widetilde{L}\delta^0 + \mathbb{E}\left[G^0\right]}{\alpha\varepsilon^2} + \frac{1}{n}\sum_{i=1}^n \frac{\left(\widetilde{L}\delta^0 + \mathbb{E}\left[G^0\right]\right)\left(\overline{L}_i(\delta^0 + \delta_i^{\inf}) + \overline{L}_i\Delta_i^{\inf}\right)}{\alpha^3\varepsilon^4} + \frac{1}{n}\sum_{i=1}^n \frac{(\widetilde{L}\delta^0 + \mathbb{E}\left[G^0\right])\overline{L}_i\mathbb{E}\left[G^0\right]}{\alpha^2(\alpha L + \widetilde{L})\varepsilon^4}\Bigg)\end{aligned}$$

stochastic oracle calls per worker on average.

Appendix D. Variance Reduction

In this part, we modify the EF21 framework to better handle *finite-sum* problems with smooth summands. Unlike the *online/streaming case* where SGD has the optimal complexity (without additional assumption on the smoothness of stochastic trajectories) (Arjevani et al., 2023), in the *finite sum* regime, it is well-known that one can hope for convergence to the exact stationary point rather than its neighborhood. To achieve this, variance reduction techniques are instrumental. One approach is to apply a PAGE-estimator (Li et al., 2021) instead of a random minibatch applied in SGD.

We recall the notations used in this section: $P_i^t = \left\|\nabla f_i(x^t) - v_i^t\right\|^2$, $P^t = \frac{1}{n}\sum_{i=1}^n P_i^t$, $V_i^t = \left\|v_i^t - g_i^t\right\|^2$, $V^t = \frac{1}{n}\sum_{i=1}^n V_i^t$, where v_i^t is a PAGE estimator. As before, $G^t = \frac{1}{n}\sum_{i=1}^n G_i^t$, $G_i^t = \left\|\nabla f_i(x^t) - g_i^t\right\|^2$.

Lemma 11 *Let Assumption 3 hold, and let v_i^{t+1} be a PAGE estimator, i. e. for $b_i^t \sim \mathrm{Be}(p_i)$*

$$v_i^{t+1} = \begin{cases} \nabla f_i(x^{t+1}) & \text{if} \quad b_i^t = 1, \\ v_i^t + \frac{1}{\tau_i}\sum\limits_{j\in I_i^t}\left(\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)\right) & \text{if} \quad b_i^t = 0, \end{cases}$$

for all $i = 1, \ldots, n$, $t \geq 0$. Then

$$\mathbb{E}\left[P^{t+1}\right] \leq (1-p_{\min})\mathbb{E}\left[P^t\right] + \widetilde{\mathcal{L}}^2\mathbb{E}\left[\left\|x^{t+1} - x^t\right\|^2\right],$$

where $\widetilde{\mathcal{L}}^2 = \frac{1}{n}\sum_{i=1}^n \frac{(1-p_i)\mathcal{L}_i^2}{\tau_i}$, $p_{\min} = \min_{i=1,\ldots,n} p_i$, and $P_i^t = \left\|\nabla f_i(x^t) - v_i^t\right\|^2$, $P^t = \frac{1}{n}\sum_{i=1}^n P_i^t$.

Proof

$$\begin{aligned}
\mathbb{E}\left[P_i^{t+1}\right] &= \mathbb{E}\left[\left\|v_i^{t+1} - \nabla f_i(x^{t+1})\right\|^2\right] \\
&= (1-p_i)\mathbb{E}\left[\left\|v_i^t + \frac{1}{\tau_i}\sum_{j\in I_i^t}(\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)) - \nabla f_i(x^{t+1})\right\|^2\right] \\
&= (1-p_i)\mathbb{E}\left[\left\|v_i^t - \nabla f_i(x^t) + \widetilde{\Delta}_i^t - \nabla f_i(x^{t+1}) + \nabla f_i(x^t)\right\|^2\right] \\
&= (1-p_i)\mathbb{E}\left[\left\|v_i^t - \nabla f_i(x^t) + \widetilde{\Delta}_i^t - \Delta_i^t\right\|^2\right] \\
&\overset{(i)}{=} (1-p_i)\mathbb{E}\left[\left\|v_i^t - \nabla f_i(x^t)\right\|^2\right] + (1-p_i)\mathbb{E}\left[\left\|\widetilde{\Delta}_i^t - \Delta_i^t\right\|^2\right] \\
&\overset{(ii)}{\leq} (1-p_i)\mathbb{E}\left[P_i^t\right] + \frac{(1-p_i)\mathcal{L}_i^2}{\tau_i}\mathbb{E}\left[\left\|x^{t+1} - x^t\right\|^2\right] \\
&\leq (1-p_{\min})\mathbb{E}\left[P_i^t\right] + \frac{(1-p_i)\mathcal{L}_i^2}{\tau_i}\mathbb{E}\left[\left\|x^{t+1} - x^t\right\|^2\right],
\end{aligned}$$

where equality (i) holds because $\mathbb{E}\left[\widetilde{\Delta}_i^t - \Delta_i^t \mid x^t, x^{t+1}, v_i^t\right] = 0$, and (ii) holds by Assumption 3.

It remains to average the above inequality over $i = 1, \ldots, n$. ■

Lemma 12 *Let Assumptions 1 and 3 hold, let v_i^{t+1} be a PAGE estimator, i. e. for $b_i^t \sim \mathrm{Be}(p_i)$ and for all $i = 1, \ldots, n$, $t \geq 0$*

$$v_i^{t+1} = \begin{cases} \nabla f_i(x^{t+1}) & \text{if} \quad b_i^t = 1, \\ v_i^t + \frac{1}{\tau_i} \sum\limits_{j \in I_i^t} \left(\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)\right) & \text{if} \quad b_i^t = 0, \end{cases} \tag{23}$$

and let g_i^{t+1} be an EF21 estimator, i. e.

$$g_i^{t+1} = g_i^t + \mathcal{C}(v_i^{t+1} - g_i^t), \quad g_i^0 = \mathcal{C}\left(v_i^0\right)$$

for all $i = 1, \ldots, n$, $t \geq 0$. Then

$$\mathbb{E}\left[V^{t+1}\right] \leq (1-\theta)\mathbb{E}\left[V^t\right] + 2\beta p_{\max}\mathbb{E}\left[P^t\right] + \beta\left(2\widetilde{L}^2 + \widetilde{\mathcal{L}}^2\right)\mathbb{E}\left[\left\|x^{t+1} - x^t\right\|^2\right], \tag{24}$$

where $\widetilde{\mathcal{L}} = \frac{1}{n}\sum_{i=1}^n \frac{(1-p_i)\mathcal{L}_i^2}{\tau_i}$, $p_{\max} = \max_{i=1,\ldots,n} p_i$, $\theta = 1-(1-\alpha)(1+s)$, $\beta = (1-\alpha)\left(1+s^{-1}\right)$ for any $s > 0$, and $P_i^t = \left\|\nabla f_i(x^t) - v_i^t\right\|^2$, $P^t = \frac{1}{n}\sum_{i=1}^n P_i^t$, $V_i^t = \left\|v_i^t - g_i^t\right\|^2$, $V^t = \frac{1}{n}\sum_{i=1}^n V_i^t$.

Proof Following the steps in proof of Lemma 3, but with $\nabla f_i(x^{t+1})$ and $\nabla f_i(x^t)$ being substituted by their estimators v_i^{t+1} and v_i^t, we end up with an analogue of (14)

$$\mathbb{E}\left[\left\|g_i^{t+1} - v_i^{t+1}\right\|^2\right] \quad \leq \quad (1-\theta)\mathbb{E}\left[\left\|g_i^t - v_i^t\right\|^2\right] + \beta\mathbb{E}\left[\left\|v_i^{t+1} - v_i^t\right\|^2\right], \tag{25}$$

where $\theta = 1-(1-\alpha)(1+s)$, $\beta = (1-\alpha)\left(1+s^{-1}\right)$ for any $s > 0$. Then

$$\begin{aligned}
\mathbb{E}\left[V_i^t\right] &= \mathbb{E}\left[\left\|g_i^{t+1} - v_i^{t+1}\right\|^2\right] \\
&\overset{(25)}{\leq} (1-\theta)\mathbb{E}\left[\left\|g_i^t - v_i^t\right\|^2\right] + \beta\mathbb{E}\left[\left\|v_i^{t+1} - v_i^t\right\|^2\right] \\
&= (1-\theta)\mathbb{E}\left[\left\|g_i^t - v_i^t\right\|^2\right] + \beta\mathbb{E}\left[\mathbb{E}\left[\left\|v_i^{t+1} - v_i^t\right\|^2 \mid v_i^t, x^t, x^{t+1}\right]\right] \\
&\overset{(i)}{=} (1-\theta)\mathbb{E}\left[V_i^t\right] + \beta p_i\mathbb{E}\left[\left\|v_i^t - \nabla f_i(x^{t+1})\right\|^2\right] \\
&\qquad + \beta(1-p_i)\mathbb{E}\left[\left\|\frac{1}{\tau_i}\sum_{j\in I_i^t}\left(\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)\right)\right\|^2\right] \\
&= (1-\theta)\mathbb{E}\left[V_i^t\right] + \beta p_i\mathbb{E}\left[\left\|v_i^t - \nabla f_i(x^{t+1})\right\|^2\right] + \beta(1-p_i)\mathbb{E}\left[\left\|\widetilde{\Delta}_i^t\right\|^2\right] = (*).
\end{aligned}$$

where in (i) we use the definition of PAGE estimator (23). Next, we continue by using Young's inequality (42) with $s=1$ in (ii)

$$
\begin{aligned}
(*) \quad &\overset{(ii)}{=} (1-\theta)\mathbb{E}\left[V_i^t\right] + 2\beta p_i \mathbb{E}\left[\left\|v_i^t - \nabla f_i(x^t)\right\|^2\right] \\
&\qquad +2\beta p_i \mathbb{E}\left[\left\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right\|^2\right] + \beta(1-p_i)\mathbb{E}\left[\left\|\widetilde{\Delta}_i^t\right\|^2\right] \\
&= (1-\theta)\mathbb{E}\left[V_i^t\right] + 2\beta p_i \mathbb{E}\left[P_i^t\right] + 2\beta p_i \mathbb{E}\left[\left\|\Delta_i^t\right\|^2\right] + \beta(1-p_i)\mathbb{E}\left[\left\|\widetilde{\Delta}_i^t\right\|^2\right] \\
&\overset{(iii)}{=} (1-\theta)\mathbb{E}\left[V_i^t\right] + 2\beta p_i \mathbb{E}\left[P_i^t\right] + \beta(2p_i+1-p_i)\mathbb{E}\left[\left\|\Delta_i^t\right\|^2\right] \\
&\qquad +\beta(1-p_i)\mathbb{E}\left[\left\|\widetilde{\Delta}_i^t - \Delta_i^t\right\|^2\right] \\
&\overset{(iv)}{\leq} (1-\theta)\mathbb{E}\left[V_i^t\right] + 2\beta p_i \mathbb{E}\left[P_i^t\right] + \beta(1+p_i)L_i^2\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right] \\
&\qquad +\beta\frac{(1-p_i)\mathcal{L}_i^2}{\tau_i}\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right] \\
&\leq (1-\theta)\mathbb{E}\left[V_i^t\right] + 2\beta p_{\max} \mathbb{E}\left[P_i^t\right] + \beta\left(2L_i^2 + \frac{(1-p_i)\mathcal{L}_i^2}{\tau_i}\right)\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right],
\end{aligned}
$$

where (iii) is due to bias-variance decomposition, (iv) makes use of Assumptions 1 and 3, and the last step is due to $p_i \leq 1$, $p_i \leq p_{\max}$. It remains to average the above inequality over $i=1,\ldots,n$.

■

Theorem 13 *Let Assumptions 1 and 3 hold, and let the stepsize in Algorithm 2 be set as*

$$
0 < \gamma \leq \left(L + \sqrt{\frac{4\beta}{\theta}\widetilde{L}^2 + 2\left(\frac{3\beta}{\theta}\frac{p_{\max}}{p_{\min}} + \frac{1}{p_{\min}}\right)\widetilde{\mathcal{L}}^2}\right)^{-1}. \tag{26}
$$

Fix $T \geq 1$ and let $\hat{x}^T$ be chosen from the iterates $x^0, x^1, \ldots, x^{T-1}$ uniformly at random. Then

$$
\mathbb{E}\left[\left\|\nabla f(\hat{x}^T)\right\|^2\right] \leq \frac{2\Psi^0}{\gamma T}, \tag{27}
$$

where $\Psi^t \overset{def}{=} f(x^t) - f^{\inf} + \frac{\gamma}{\theta}V^t + \frac{\gamma}{p_{\min}}\left(1 + \frac{2\beta p_{\min}}{\theta}\right)P^t$, $p_{\max} = \max_{i=1,\ldots,n} p_i$, $p_{\min} = \min_{i=1,\ldots,n} p_i$, $\widetilde{L} = \sqrt{\frac{1}{n}\sum_{i=1}^n L_i^2}$, $\theta = 1-(1-\alpha)(1+s)$, $\beta = (1-\alpha)\left(1+s^{-1}\right)$ for any $s>0$.

Proof We apply Lemma 21 and split the error $\left\|g_i^t - \nabla f_i(x^t)\right\|^2$ in two parts

$$\begin{aligned} f(x^{t+1}) &\leq f(x^t) - \frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right) R^t + \frac{\gamma}{2}\left\|g^t - \nabla f(x^t)\right\|^2 \\ &\leq f(x^t) - \frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right) R^t \\ &\quad + \gamma\left\|g^t - v^t\right\|^2 + \gamma\mathbb{E}\left[\left\|v^t - \nabla f(x^t)\right\|^2\right] \\ &\leq f(x^t) - \frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right) R^t \\ &\quad + \gamma\frac{1}{n}\sum_{i=1}^n \left\|g_i^t - v_i^t\right\|^2 + \gamma\frac{1}{n}\sum_{i=1}^n \left\|v_i^t - \nabla f_i(x^t)\right\|^2 \\ &= f(x^t) - \frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right) R^t + \gamma V^t + \gamma P^t, \end{aligned} \tag{28}$$

where we used notation $R^t = \left\|\gamma g^t\right\|^2 = \left\|x^{t+1} - x^t\right\|^2$, and applied (41) and (42).

Subtracting $f^{\inf}$ from both sides of the above inequality, taking expectation and using the notation $\delta^t = f(x^{t+1}) - f^{\inf}$, we get

$$\mathbb{E}\left[\delta^{t+1}\right] \leq \mathbb{E}\left[\delta^t\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\mathbb{E}\left[R^t\right] + \gamma\mathbb{E}\left[V^t\right] + \gamma\mathbb{E}\left[P^t\right]. \tag{29}$$

Further, Lemma 11 and 12 provide the recursive bounds for the last two terms of (29)

$$\mathbb{E}\left[P^{t+1}\right] \leq (1 - p_{\min})\mathbb{E}\left[P^t\right] + \widetilde{\mathcal{L}}^2\mathbb{E}\left[R_t\right], \tag{30}$$

$$\mathbb{E}\left[V^{t+1}\right] \leq (1-\theta)\mathbb{E}\left[V^t\right] + \beta\left(2\widetilde{L}^2 + \widetilde{\mathcal{L}}^2\right)\mathbb{E}\left[R_t\right] + 2\beta p_{\max}\mathbb{E}\left[P^t\right]. \tag{31}$$

Adding (29) with a $\frac{\gamma}{\theta}$ multiple of (31) we obtain

$$\begin{aligned} \mathbb{E}\left[\delta^{t+1}\right] + \frac{\gamma}{\theta}\mathbb{E}\left[V^{t+1}\right] &\leq \mathbb{E}\left[\delta^t\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f\left(x^t\right)\right\|^2\right] - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\mathbb{E}\left[R^t\right] + \gamma\mathbb{E}\left[V^t\right] \\ &\quad + \gamma\mathbb{E}\left[P^t\right] + \frac{\gamma}{\theta}\left((1-\theta)\,\mathbb{E}\left[V^t\right] + Ar^t + C\mathbb{E}\left[P^t\right]\right) \\ &\leq \delta^t + \frac{\gamma}{\theta}\mathbb{E}\left[V^t\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f\left(x^t\right)\right\|^2\right] - \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{\gamma A}{\theta}\right)\mathbb{E}\left[R^t\right] \\ &\quad + \gamma\left(1 + \frac{C}{\theta}\right)\mathbb{E}\left[P^t\right], \end{aligned}$$

where we denote $A \overset{\text{def}}{=} \beta\left(2\widetilde{L}^2 + \widetilde{\mathcal{L}}^2\right)$, $C \overset{\text{def}}{=} 2\beta p_{\max}$.

Then adding the above inequality with a $\frac{\gamma}{p_{\min}}\left(1 + \frac{C}{\theta}\right)$ multiple of (30), we get

$$
\begin{aligned}
\mathbb{E}\left[\Phi^{t+1}\right] &= \mathbb{E}\left[\delta^{t+1}\right] + \frac{\gamma}{\theta}\mathbb{E}\left[V^{t+1}\right] + \frac{\gamma}{p_{\min}}\left(1+\frac{C}{\theta}\right)\mathbb{E}\left[P^{t+1}\right] \\
&\leq \delta^t + \frac{\gamma}{\theta}\mathbb{E}\left[V^t\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f\left(x^t\right)\right\|^2\right] - \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{\gamma A}{\theta}\right)\mathbb{E}\left[R^t\right] \\
&\quad + \gamma\left(1+\frac{C}{\theta}\right)\mathbb{E}\left[P^t\right] \\
&\quad + \frac{\gamma}{p_{\min}}\left(1+\frac{C}{\theta}\right)\left((1-p_{\min})\mathbb{E}\left[P^t\right] + \widetilde{\mathcal{L}}^2\mathbb{E}\left[R^t\right]\right) \\
&\leq \mathbb{E}\left[\delta^t\right] + \frac{\gamma}{\theta}\mathbb{E}\left[V^t\right] + \frac{\gamma}{p_{\min}}\left(1+\frac{C}{\theta}\right)\mathbb{E}\left[P^t\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f\left(x^t\right)\right\|^2\right] \\
&\quad - \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{\gamma A}{\theta} - \frac{\gamma}{p_{\min}}\left(1+\frac{C}{\theta}\right)\widetilde{\mathcal{L}}^2\right)\mathbb{E}\left[R^t\right] \\
&= \mathbb{E}\left[\Phi^t\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f\left(x^t\right)\right\|^2\right] \\
&\quad - \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{\gamma A}{\theta} - \frac{\gamma}{p_{\min}}\left(1+\frac{C}{\theta}\right)\widetilde{\mathcal{L}}^2\right)\mathbb{E}\left[R^t\right]. \qquad (32)
\end{aligned}
$$

The coefficient in front of $\mathbb{E}\left[R^t\right]$ simplifies after substitution by A and C

$$
\frac{\gamma A}{\theta} + \frac{\gamma}{p_{\min}}\left(1+\frac{C}{\theta}\right)\widetilde{\mathcal{L}}^2 \leq \frac{2\beta}{\theta}\widetilde{L}^2 + \left(\frac{3\beta}{\theta}\frac{p_{\max}}{p_{\min}} + \frac{1}{p_{\min}}\right)\widetilde{\mathcal{L}}^2.
$$

Thus by Lemma 20 and the stepsize choice, the last term in (32) is not positive. By summing up inequalities for $t = 0, \ldots, T-1$, and rearranging we get (27).

■

Corollary 14 *Let assumptions of Theorem 13 hold,*

$$
\begin{aligned}
v_i^0 &= g_i^0 = \nabla f_i(x^0), \quad i = 1, \ldots, n, \\
\gamma &= \left(L + \sqrt{\frac{4\beta}{\theta}\widetilde{L}^2 + 2\left(\frac{3\beta}{\theta}\frac{p_{\max}}{p_{\min}} + \frac{1}{p_{\min}}\right)\widetilde{\mathcal{L}}^2}\right)^{-1}, \\
p_i &= \frac{\tau_i}{\tau_i + m}, \quad i = 1, \ldots, n.
\end{aligned}
$$

Then, after T iterations/communication rounds of EF21-PAGE *we have $\mathbb{E}\left[\left\|\nabla f(\hat{x}^T)\right\|^2\right] \leq \varepsilon^2$. It requires*

$$
T = \mathcal{O}\left(\frac{(\widetilde{L} + \widetilde{\mathcal{L}})\delta^0}{\alpha\varepsilon^2}\sqrt{\frac{p_{\max}}{p_{\min}}} + \frac{\sqrt{m_{\max}}\widetilde{\mathcal{L}}\delta^0}{\varepsilon^2}\right)
$$

iterations/communications rounds,

$$\#grad_i \quad = \quad \mathcal{O}\left(m + \frac{\tau_i(\widetilde{L}+\widetilde{\mathcal{L}})\delta^0}{\alpha\varepsilon^2}\sqrt{\frac{p_{\max}}{p_{\min}}} + \frac{\tau_i\sqrt{m}\widetilde{\mathcal{L}}\delta^0}{\varepsilon^2}\right)$$

stochastic oracle calls for worker i, and

$$\overline{\#grad} \quad = \quad \mathcal{O}\left(m + \frac{\tau(\widetilde{L}+\widetilde{\mathcal{L}})\delta^0}{\alpha\varepsilon^2}\sqrt{\frac{p_{\max}}{p_{\min}}} + \frac{\tau\sqrt{m}\widetilde{\mathcal{L}}\delta^0}{\varepsilon^2}\right)$$

stochastic oracle calls per worker on average, where $\tau = \frac{1}{n}\sum_{i=1}^n \tau_i$, $p_{\max} = \max_{i=1,\ldots,n} p_i$, $p_{\min} = \min_{i=1,\ldots,n} p_i$.

Proof The proof is straightforward using Lemma 22 and the formula: $\#\text{grad}_i = m+T\left(p_i m + (1-p_i)\tau_i\right)$. ■

Appendix E. Partial Participation

In this section, we further motivate the option for partial participation of the clients – a feature important in federated learning. Later, we continue with a rigorous proof of EF21-PP algorithm.

Most of the works in compressed distributed optimization deal with full worker participation, i.e., the case when all clients are involved in computation and communication at every iteration. However, in the practice of federated learning, only a subset of clients are allowed to participate at each training round. This limitation comes mainly due to the following two reasons. First, clients (e.g., mobile devices) may wish to join or leave the network randomly. Second, it is often prohibitive to wait for all available clients since stragglers can significantly slow down the training process. Although many existing works (Gorbunov et al., 2021; Horváth and Richtárik, 2021; Philippenko and Dieuleveut, 2020; Karimireddy et al., 2020; Yang et al., 2021; Cho et al., 2020) allow for partial participation, they assume either unbiased compressors or no compression at all.

We provide a simple analysis of partial participation, which works with *biased compressors* and builds upon the EF21 mechanism.

Below we list the statements of key intermediate results, the formal proof is deferred the extended version of our paper (Fatkhullin et al., 2025) due to space limitations.

Lemma 15 *For Algorithm 3 it holds*

$$\mathbb{E}\left[G^{t+1}\right] \leq (1-\theta_p)\mathbb{E}\left[G^t\right] + B\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right] \tag{33}$$

with $\theta_p \overset{def}{=} \rho p_{min} + \theta p_{max} - \rho - (p_{max}-p_{min})$, $B \overset{def}{=} \frac{1}{n}\sum_{i=1}^n \left(\beta p_i + \left(1+\rho^{-1}\right)(1-p_i)\right) L_i^2$, $p_{max} \overset{def}{=} \max_{1\leq i\leq n} p_i$, $p_{min} \overset{def}{=} \min_{1\leq i\leq n} p_i$, $\theta = 1-(1+s)(1-\alpha)$, $\beta = \left(1+\frac{1}{s}\right)(1-\alpha)$ *and small enough* $\rho, s > 0$.

Lemma 16 *[To simplify the rates for partial participation] Let* B *and* θ_p *be defined as in Theorem 17, and let* $p_i = p > 0$ *for all* $i = 1, \ldots, n$ *. Then there exist* $\rho, s > 0$ *such that*

$$\theta_p \geq \frac{p\alpha}{2}, \tag{34}$$

$$0 < \frac{B}{\theta_p} \leq \left(\frac{4\widetilde{L}}{p\alpha}\right)^2. \tag{35}$$

Theorem 17 *Let Assumption 1 hold, and let the stepsize in Algorithm 3 be set as*

$$0 < \gamma \leq \left(L + \sqrt{\frac{B}{\theta_p}}\right)^{-1}. \tag{36}$$

Fix $T \geq 1$ *and let* $\hat{x}^T$ *be chosen from the iterates* $x^0, x^1, \ldots, x^{T-1}$ *uniformly at random. Then*

$$\mathbb{E}\left[\left\|\nabla f(\hat{x}^T)\right\|^2\right] \leq \frac{2\left(f(x^0)-f^{\inf}\right)}{\gamma T} + \frac{\mathbb{E}\left[G^0\right]}{\theta_p T} \tag{37}$$

with $\theta_p = \rho p_{min} + \theta p_{max} - \rho - (p_{max}-p_{min})$, $B = \frac{1}{n}\sum_{i=1}^n \left(\beta p_i + \left(1+\rho^{-1}\right)(1-p_i)\right) L_i^2$, $p_{max} = \max_{1\leq i\leq n} p_i$, $p_{min} = \min_{1\leq i\leq n} p_i$, $\theta = 1-(1+s)(1-\alpha)$, $\beta = \left(1+\frac{1}{s}\right)(1-\alpha)$ *and* $\rho, s > 0$.

Corollary 18 *Let assumptions of Theorem 17 hold,*

$$\begin{aligned} g_i^0 &= \nabla f_i(x^0), \qquad i = 1, \ldots, n, \\ \gamma &= \left(L + \sqrt{\frac{B}{\theta_p}}\right)^{-1}, \\ p_i &= p, \qquad i = 1, \ldots, n, \end{aligned}$$

where B and θ_p are given in Theorem 17. Then, after T iterations/communication rounds of EF21-PP *we have $\mathbb{E}\left[\left\|\nabla f(\hat{x}^T)\right\|^2\right] \leq \varepsilon^2$. It requires*

$$T = \#grad = \mathcal{O}\left(\frac{\widetilde{L}\delta^0}{p\alpha\varepsilon^2}\right)$$

iterations/communications rounds/gradint computations at each node.

Appendix F. Bidirectional Compression

The main idea of the proof is to split the deviation error coming from worker's compressor and the server's compressor. That is we need to control the terms

$$\left\|g^t - \widetilde{g}^t\right\|^2 \qquad \text{and} \qquad \left\|\widetilde{g}^t - \nabla f(x^t)\right\|^2,$$

where

$$g^{t+1} = g^t + \mathcal{C}_M(\widetilde{g}^{t+1} - g^t) \qquad \text{and} \quad \widetilde{g}_i^{t+1} = \widetilde{g}_i^t + \mathcal{C}_w(\nabla f_i(x^{t+1}) - \widetilde{g}_i^t), \quad \widetilde{g}^{t+1} = \frac{1}{n}\sum_{i=1}^n \widetilde{g}_i^{t+1}.$$

This is conceptually similar to the proof strategy for variance reduction extension, but the source of the second deviation error in this case is different and comes from server level compression rather than sampling stochastic gradients.

For the formal proof of this result, please refer to the extended version of this work (Fatkhullin et al., 2025).

Appendix G. Heavy Ball Momentum

In this section, we study the momentum version of EF21. In particular, we focus on Polyak style momentum (Polyak, 1964; Yang et al., 2016). Let g^t be a gradient estimator at iteration t and v^t is some vector, then the update rule of *heavy ball* (HB) can be written as

$$\begin{cases} x^{t+1} = x^t - \gamma v^t \\ v^{t+1} = \eta v^t + g^{t+1}, \end{cases}$$

where $\eta \in [0, 1)$ is the *momentum parameter*, and $\gamma > 0$ is the stepsize. To combine this algorithm with EF21, we use EF21 estimator to approximate g^t. The formal pseudocode in distributed setting is presented in Algorithm 5.

We defer the convergence proof of EF21-HB to the extended version of this work (Fatkhullin et al., 2025) due to space limitations.

Appendix H. Composite Setting

Now we focus on solving a composite optimization problem

$$\min_{x\in\mathbb{R}^d} \Phi(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n f_i(x) + r(x), \tag{38}$$

where each $f_i(\cdot)$ is L_i-smooth (possibly non-convex), $r(\cdot)$ is convex, and $\Phi^{\inf} = \inf_{x\in\mathbb{R}^d}\Phi(x) > -\infty$. This is a standard and important generalization of problem (1). In particular, it includes optimization problems over convex compact sets and l_1-regularization (LASSO).

For any $\gamma > 0$, $x \in \mathbb{R}^d$, recall that the proximal mapping of function $r(\cdot)$ (prox-operator) is defined as

$$\text{prox}_{\gamma r}(x) = \arg\min_{y\in\mathbb{R}^d}\left\{r(y) + \frac{1}{2\gamma}\|y-x\|^2\right\}. \tag{39}$$

To evaluate convergence in composite case, we define the *generalized gradient mapping* at a point $x \in \mathbb{R}^d$ with a parameter γ

$$\mathcal{G}_\gamma(x) \stackrel{\text{def}}{=} \frac{1}{\gamma}\left(x - \text{prox}_{\gamma r}(x - \gamma\nabla f(x))\right).$$

One can verify that the above quantity is a well-defined evaluation metric (Beck, 2017). Namely, for any $x^* \in \mathbb{R}^d$, it holds that $\mathcal{G}_\gamma(x) = 0$ if and only if x^* is a stationary point of (38), and in a special case when $r \equiv 0$, we have $\mathcal{G}_\gamma(x) = \nabla f(x)$.

When there is no compression, the convergence analysis of proximal gradient descent (see, e.g., Section 10.3 in (Beck, 2017)) consists in showing a descent lemma with repect to the squared norm of gradient mapping, i.e., for any x^t

$$\Phi(x^t) - \Phi(x^{t+1}) \geq \gamma\left(1 - \gamma L/2\right)\|\mathcal{G}_\gamma(x)\|^2.$$

However, when there is a non-trivial compression, such inequality may not hold. The main idea of the analysis below is to upper bound the squared norm of gradient mapping $\|\mathcal{G}_\gamma(x)\|^2$ with certain error terms proportional to $\left\|x^{t+1} - x^t\right\|^2$ and $\left\|g^t - \nabla f(x^t)\right\|^2$, which can be controlled using recursions from EF21 analysis (Lemma 3).

We defer the formal proof of this result to the extended version of this paper (Fatkhullin et al., 2025).

Appendix I. Useful Lemma

Lemma 19 (Basic Facts) *For all $a, b, x_1, \ldots, x_n \in \mathbb{R}^d$, $s > 0$ and $p \in (0, 1]$ the following inequalities hold*

$$\begin{aligned}
\langle a, b\rangle &\leq \frac{\|a\|^2}{2s} + \frac{s\|b\|^2}{2}, && (40)\\
\|a+b\|^2 &\leq (1+s)\|a\|^2 + (1+1/s)\|b\|^2, && (41)\\
\left\|\frac{1}{n}\sum_{i=1}^n x_i\right\|^2 &\leq \frac{1}{n}\sum_{i=1}^n \|x_i\|^2, && (42)\\
\left(1-\frac{p}{2}\right)^{-1} &\leq 1+p, && (43)\\
\left(1+\frac{p}{2}\right)(1-p) &\leq 1-\frac{p}{2}, && (44)\\
\log(1-p) &\leq -p. && (45)
\end{aligned}$$

Lemma 20 (Lemma 5 of (Richtárik et al., 2021)) *If $0 \leq \gamma \leq \frac{1}{\sqrt{a}+b}$, then $a\gamma^2 + b\gamma \leq 1$. Moreover, the bound is tight up to the factor of 2 since $\frac{1}{\sqrt{a}+b} \leq \min\left\{\frac{1}{\sqrt{a}}, \frac{1}{b}\right\} \leq \frac{2}{\sqrt{a}+b}$.*

Lemma 21 (Lemma 2 of (Li et al., 2021)) *Suppose that function f is L-smooth and let $x^{t+1} \overset{def}{=} x^t - \gamma g^t$, where $g^t \in \mathbb{R}^d$ is any vector, and $\gamma > 0$ any scalar. Then we have*

$$f(x^{t+1}) \leq f(x^t) - \frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\left\|x^{t+1} - x^t\right\|^2 + \frac{\gamma}{2}\left\|g^t - \nabla f(x^t)\right\|^2.$$

Lemma 22 (Lemma 3 of (Richtárik et al., 2021)) *Let $0 < \alpha < 1$ and for $s > 0$ let $\theta(s)$ and $\beta(s)$ be defined as*

$$\theta(s) \overset{def}{=} 1 - (1-\alpha)(1+s), \qquad \beta(s) \overset{def}{=} (1-\alpha)(1+s^{-1}).$$

Then the solution of the optimization problem

$$\min_s \left\{\frac{\beta(s)}{\theta(s)} \;:\; 0 < s < \frac{\alpha}{1-\alpha}\right\}$$

is given by $s^* = \frac{1}{\sqrt{1-\alpha}} - 1$. Furthermore, $\theta(s^*) = 1 - \sqrt{1-\alpha}$, $\beta(s^*) = \frac{1-\alpha}{1-\sqrt{1-\alpha}}$ and

$$\sqrt{\frac{\beta(s^*)}{\theta(s^*)}} = \frac{1}{\sqrt{1-\alpha}} - 1 = \frac{1}{\alpha} + \frac{\sqrt{1-\alpha}}{\alpha} - 1 \leq \frac{2}{\alpha} - 1.$$

In the trivial case $\alpha = 1$, we have $\frac{\beta(s)}{\theta(s)} = 0$ for any $s > 0$, and above inequality is satisfied.

Lemma 23 *Let (arbitrary scalar) non-negative sequences $\{s^t\}_{t\geq 0}$, and $\{r^t\}_{t\geq 0}$ satisfy*

$$\sum_{t=0}^{T-1} s^{t+1} \leq (1-\theta)\sum_{t=0}^{T-1} s^t + C\sum_{t=0}^{T-1} r^t$$

for some parameters $\theta \in (0, 1]$, $C > 0$. *Then for all* $T \geq 0$

$$\sum_{t=0}^{T-1} s^t \leq \frac{s^0}{\theta} + \frac{C}{\theta} \sum_{t=0}^{T-1} r^t.$$

Proof The proof follows immediately by canceling out the common terms on both sides and then dividing by $\theta > 0$. ■

Appendix J. Extra Experiments

In this section, we give missing details on the experiments from Section 6, and provide additional experiments.

J.1 Non-Convex Logistic Regression: Additional Experiments and Details

Datasets, hardware and implementation. We use standard LibSVM data sets (Chang and Lin, 2011), and split each data set among n clients. For experiments 1, 3, 4 and 5, we chose $n = 20$ whereas for the experiment 2 we consider $n = 100$. The first $n - 1$ clients own equal parts, and the remaining part, of size $N - n \cdot \lfloor N/n \rfloor$, is assigned to the last client. We consider the heterogeneous data distribution regime (i.e. we do not make any additional assumptions on data similarity between workers). A summary of data sets and details of splitting data among workers can be found in Tables 4 and 6. The algorithms are implemented in Python 3.8; we use 3 different CPU cluster node types in all experiments: 1) AMD EPYC 7702 64-Core; 2) Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz; 3) Intel(R) Xeon(R) Gold 6248 CPU @ 2.50GHz. In all algorithms involving compression, we use Top-k (Alistarh et al., 2017) as a canonical example of contractive compressor $\mathcal{C}$, and fix the compression ratio $k/d \approx 0.01$, where d is the number of features in the data set. For all algorithms, at each iteration we compute the squared norm of the exact/full gradient for comparison of the methods performance. We terminate our algorithms either if they reach the certain number of iterations or the following stopping criterion is satisfied: $\left\|\nabla f(x^t)\right\|^2 \leq 10^{-7}$.

In all experiments, the stepsize is set to the largest stepsize predicted by theory for EF21 multiplied by some constant multiplier which was individually tuned in all cases.

Table 4: Summary of the data sets and splitting of the data among clients for Experiments 1, 3, 4, and 5. Here N_i denotes the number of datapoints per client.

Data set	n	N (total # of datapoints)	d (# of features)	k	N_i
mushrooms	20	8,120	112	2	406
w8a	20	49,749	300	2	2,487
a9a	20	32,560	123	2	1,628
phishing	20	11,055	68	1	552
real-sim	20	72,309	20,958	210	3615

Experiment 1: Fast convergence with variance reductions (extra details). The parameters p_i of the PAGE estimator are set to $p_i = p \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \frac{\tau_i}{\tau_i + N_i}$, where τ_i is the batchsize for clients $i = 1, \ldots, n$ (see Table 5 for details). In our experiments, we assume that the sampling of Bernoulli random variable is performed on server side (which means that at each iteration for all clients $b_i^t = 1$ or $b_i^t = 0$). And if $b_i^t = 0$, then in line 5 of Algorithm 2 I_i^t is sampled without replacement uniformly at random. Table 5 shows the selection of parameter p for each experiment.

For each batchsize from the set[11]

$$\{95\%, 50\%, 25\%, 12.5\%, 6.5\%, 3\%\},$$

11. By $50\%, 25\%$ (and so on) we refer to a batchsize, which is equals to $\lfloor 0.5N_i \rfloor$, $\lfloor 0.25N_i \rfloor$ (and so on) for all clients $i = 1, \ldots, n$.

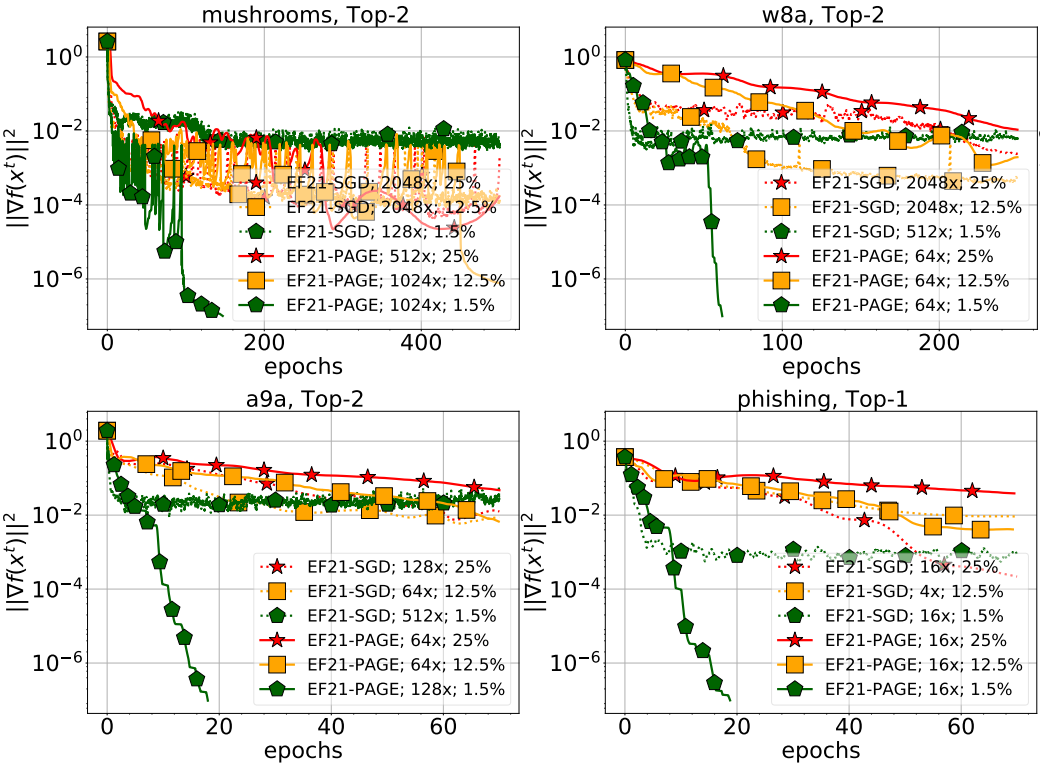


Figure 3: Comparison of EF21-PAGE and EF21-SGD with tuned step-sizes. By $1\times, 2\times, 4\times$ (and so on) we indicate that the stepsize was set to a multiple of the largest stepsize predicted by theory for EF21. By 25%, 12.5% and 1.5% we refer to batch-sizes equal $\lfloor 0.25N_i \rfloor$, $\lfloor 0.125N_i \rfloor$ and $\lfloor 0.015N_i \rfloor$ for all clients $i = 1, \ldots, n$, where N_i denotes the size of local data set.

we tune the stepsize multiplier for EF21-PAGE within the set

$$\{0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048\}.$$

The best pair (batchsize, stepsize multiplier) is chosen in such a way that it gives the best convergence in terms of $\#\text{bits}/n(C \to S)$. In the rest of the experiments, fine tuning is performed in a similar fashion.

Table 5: Summary of the parameter choice of p.

Data set	25%	12.5%	1.5%
`mushrooms`	0.1992	0.1097	0.0146
`w8a`	0.1998	0.1108	0.0147
`a9a`	0.2	0.1109	0.0145
`phishing`	0.2	0.1111	0.0143
`real-sim`	0.1999	0.1109	0.0147

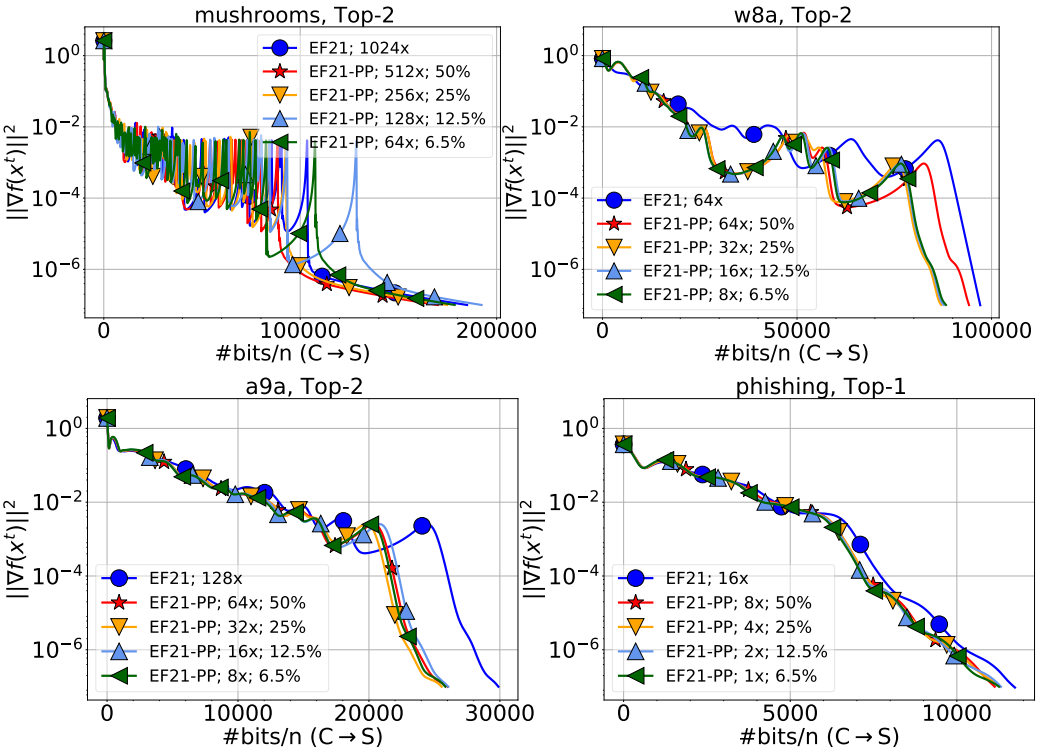


Figure 4: Comparison of EF21-PP and EF21 with tuned step-sizes. By $1\times, 2\times, 4\times$ (and so on) we indicate that the stepsize was set to a multiple of the largest stepsize predicted by theory for EF21. By 50%, 25% , 12.5% and 6.5% we refer to a number of participating clients equal to $\lfloor 0.5n \rfloor$, $\lfloor 0.25n \rfloor$, $\lfloor 0.125n \rfloor$ and $\lfloor 0.065n \rfloor$.

Experiment 2: On the effect of partial participation of clients (extra details). In this experiment, we consider $n = 100$ and, therefore, a different data partitioning, see Table 6 for the summary.

Table 6: Summary of the data sets and splitting of the data among clients for Experiment 5. Here N_i denotes the number of datapoints per client.

Data set	n	N (total # of datapoints)	d (# of features)	k	N_i
`mushrooms`	100	8,120	112	2	81
`w8a`	100	49,749	300	2	497
`a9a`	100	32,560	123	2	325
`phishing`	100	11,055	68	1	110

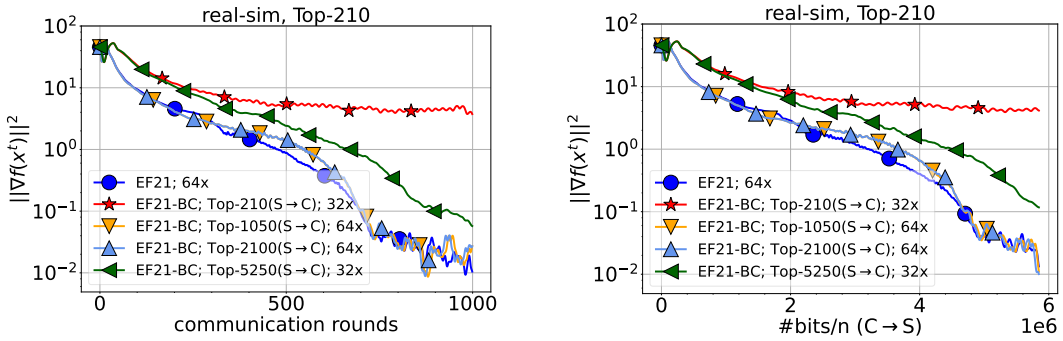
We tune the stepsize multiplier for EF21-PP within the following set:

$$\{0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096\}\,.$$

Experiment 3: On the advantages of bidirectional biased compression (extra details). Our next experiment, for each parameter k in Server-Clients compression, we tune the stepsize multiplier for EF21-BC within the following set:

$$\{0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048\}\,.$$

This experiment shows that the application of the **S**erver $\to$ **C**lients compression in EF21-BC (Alg. 4) does not significantly slow down the convergence in terms of the communication rounds but requires much less bits to be transmitted. Indeed, Figure 5a illustrates that that it is sufficient to communicate only $5\% - 15\%$ of data to perform similarly to EF21 (Alg. 7).[12] Note that EF21 communicates full vectors from the **S**erver $\to$ **C**lients, and, therefore, may have slower communication at each round. In Figure 5b we take into account only the number of bits sent from clients to the server. However, if we consider the total number of bits (see Figure 2b), then EF21-BC considerably outperforms EF21 in all cases.



(a) Convergence in communication rounds.

(b) Convergence in terms of total number of bits sent from **C**lients to the **S**erver divided by n.

Figure 5: Comparison of EF21-BC and EF21 with tuned stepsizes . By $1\times, 2\times, 4\times$ (and so on) we indicate that the stepsize was set to a multiple of the largest stepsize predicted by theory for EF21 (see the Theorem 4).

References

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1709–1720, 2017.

Dan Alistarh, Torsten Hoefler, Mikael Johansson, Sarit Khirirat, Nikola Konstantinov, and Cédric Renggli. The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1200–1205. ACM, 2017.

12. The range $5\% - 15\%$ comes from the fractions k/d for each data set.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1): 165–214, 2023.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Amir Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, 2017.

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

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27, 2011.

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *arXiv preprint arXiv:2010.01243v1*, 2020.

Ilyas Fatkhullin, Igor Sokolov, Eduard Gorbunov, Zhize Li, and Peter Richtárik. EF21 with bells & whistles: Practical algorithmic extensions of modern error feedback. *arXiv preprint arXiv:2110.03294*, 2021.

Ilyas Fatkhullin, Igor Sokolov, Eduard Gorbunov, Zhize Li, and Peter Richtárik. EF21 with bells & whistles: Six algorithmic extensions of modern error feedback. *arXiv preprint arXiv:2110.03294v2*, 2025.

Eduard Gorbunov, Dmitry Kovalev, Dmitry Makarenko, and Peter Richtárik. Linearly converging error compensated SGD. In *34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Eduard Gorbunov, Konstantin Burlachenko, Zhize Li, and Peter Richtárik. MARINA: Faster nonconvex distributed learning with compression. In *International Conference on Machine Learning*, pages 3788–3798. PMLR, 2021.

Robert M Gower, Mark Schmidt, Francis Bach, and Peter Richtárik. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108(11):1968–1983, 2020.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR, 2019.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Samuel Horváth and Peter Richtárik. A better alternative to error feedback for communication-efficient distributed learning. In *9th International Conference on Learning Representations (ICLR)*, 2021.

Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*, 2019.

Samuel Horvóth, Chen-Yu Ho, Ludovit Horvath, Atal Narayan Sahu, Marco Canini, and Peter Richtárik. Natural compression for distributed deep learning. In *Mathematical and Scientific Machine Learning*, pages 129–141. PMLR, 2022.

Rustem Islamov, Xun Qian, and Peter Richtárik. Distributed second order methods with fast rates and compressed communication. In *International conference on machine learning*, pages 4617–4628. PMLR, 2021.

Peter et al Kairouz. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *36th International Conference on Machine Learning (ICML)*, 2019.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Ahmed Khaled and Peter Richtárik. Gradient descent with compressed iterates. In *NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality*, 2019.

Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020.

Sarit Khirirat, Hamid Reza Feyzmahdavian, and Mikael Johansson. Distributed learning with compressed gradients. *arXiv preprint arXiv:1806.06573*, 2018.

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

Anastasia Koloskova, Tao Lin, S. Stich, and Martin Jaggi. Decentralized deep learning with arbitrary communication compression. In *International Conference on Learning Representations (ICLR)*, 2020.

Jakub Konečný, H. Brendan McMahan, Felix Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: strategies for improving communication efficiency. In *NIPS Private Multi-Party Machine Learning Workshop*, 2016.

Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Mathematical programming*, 171:167–215, 2018.

Guanghui Lan, Zhize Li, and Yi Zhou. A unified variance-reduced accelerated gradient method for convex optimization. In *Advances in Neural Information Processing Systems*, pages 10462–10472, 2019.

Zhize Li. ANITA: An optimal loopless accelerated variance-reduced gradient method. *arXiv preprint arXiv:2103.11333*, 2021.

Zhize Li and Jian Li. A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5569–5579, 2018.

Zhize Li and Peter Richtárik. A unified analysis of stochastic gradient methods for nonconvex federated optimization. *arXiv preprint arXiv:2006.07013*, 2020.

Zhize Li and Peter Richtárik. CANITA: Faster rates for distributed convex optimization with communication compression. *arXiv preprint arXiv:2107.09461*, 2021a.

Zhize Li and Peter Richtárik. ZeroSARAH: Efficient nonconvex finite-sum optimization with zero full gradient computation. *arXiv preprint arXiv:2103.01447*, 2021b.

Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtárik. Acceleration for compressed gradient descent in distributed and federated optimization. In *International Conference on Machine Learning (ICML)*, pages 5895–5904. PMLR, 2020.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning (ICML)*, pages 6286–6295. PMLR, 2021. arXiv:2008.10898.

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

Nicolas Loizou and Peter Richtárik. Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, 77:653–710, 2020.

Stanislaw Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2, 1963.

Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning with compressed gradient differences. *Optimization Methods and Software*, pages 1–16, 2024.

Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence O (1/kˆ 2). In *Doklady AN USSR*, volume 269, pages 543–547, 1983.

Constantin Philippenko and Aymeric Dieuleveut. Bidirectional compression in heterogeneous settings for distributed or federated learning with partial participation: tight convergence guarantees. *arXiv preprint arXiv:2006.14591*, 2020.

Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

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

Xun Qian, Peter Richtárik, and Tong Zhang. Error compensated distributed SGD can be accelerated. *arXiv preprint arXiv:2010.00091*, 2020.

Zheng Qu and Peter Richtárik. Coordinate descent with arbitrary sampling ii: Expected separable overapproximation. *Optimization Methods and Software*, 31(5):858–884, 2016.

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Mher Safaryan, Rustem Islamov, Xun Qian, and Peter Richtarik. Fednl: Making newton-type methods applicable to federated learning. In *International Conference on Machine Learning*, pages 18959–19010, 2022.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: from theory to algorithms*. Cambridge University Press, 2014.

Sebastian U. Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed updates. *Journal of Machine Learning Research*, 21 (215):1–49, 2020.

Sebastian U. Stich, J.-B. Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Hanlin Tang, Xiangru Lian, Chen Yu, Tong Zhang, and Ji Liu. DoubleSqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2020.

Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: Practical low-rank gradient compression for distributed optimization. In *Neural Information Processing Systems*, 2019.

Cong Xie, Shuai Zheng, Oluwasanmi Koyejo, Indranil Gupta, Mu Li, and Haibin Lin. CSER: Communication-efficient SGD with error reset. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12593–12603, 2020.

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

Tianbao Yang, Qihang Lin, and Zhe Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *arXiv preprint arXiv:1604.03257*, 2016.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. In *International Conference on Learning Representations*, 2020.

Haoyu Zhao, Zhize Li, and Peter Richtárik. FedPAGE: A fast local stochastic gradient method for communication-efficient federated learning. *arXiv preprint arXiv:2108.04755*, 2021.