Communication Compression for Byzantine Robust Learning: New Efficient Algorithms and Improved Rates

Ahmad Rammal KAUST École Polytechnique
Kaja Gruntkowska KAUST
Nikita Fedin MIPT
Eduard Gorbunov MBZUAI
Peter Richtárik KAUST

Abstract

Byzantine robustness is an essential feature of algorithms for certain distributed optimization problems, typically encountered in collaborative/federated learning. These problems are usually huge-scale, implying that communication compression is also imperative for their resolution. These factors have spurred recent algorithmic and theoretical developments in the literature of Byzantine-robust learning with compression. In this paper, we contribute to this research area in two main directions. First, we propose a new Byzantine-robust method with compression – Byz-DASHA-PAGE – and prove that the new method has better convergence rate (for non-convex and Polyak-Łojasiewicz smooth optimization problems), smaller neighborhood size in the heterogeneous case, and tolerates more Byzantine workers under over-parametrization than the previous method with SOTA theoretical convergence guarantees (Byz-VR-MARINA). Secondly, we develop the first Byzantine-robust method with communication compression and error feedback – Byz-EF21 – along with its bidirectional compression version – Byz-EF21-BC – and derive the convergence rates for these methods for non-convex and Polyak-Łojasiewicz smooth case. We test the proposed methods and illustrate our theoretical findings in the numerical experiments.

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1 INTRODUCTION

Contemporary machine learning and deep learning pose a number of challenges, with the ability to train increasingly complex models using datasets of enormous sizes becoming one of the most pressing issues (OpenAI, 2023). Training such models on a single machine within a reasonable timeframe is no longer feasible (Li, 2020). In response to this challenge, distributed algorithms have emerged as indispensable tools, effectively sharing the computational load across multiple machines, and hence significantly speeding up the training process. Such methods also prove invaluable when data is inherently distributed across multiple sources or locations. When this is the case, the adoption of distributed methods is not only a natural choice, but often an imperative one (Konečný et al., 2016; Kairouz et al., 2021).

While distributed learning comes with a number of benefits, it also introduces some risks. In collaborative and federated learning scenarios, these include the potential presence of *Byzantine workers*[1]. Standard methods such as Parallel Stochastic Gradient Descent (SGD) (Zinkevich et al., 2010), which rely on averaging vectors received from workers, are highly vulnerable to Byzantine attacks. Consequently, a critical need has emerged for the development and investigation of specialized methods designed to demonstrate robustness when Byzantine participants are involved, giving Byzantine-robustness significant attention in recent years (Lyu et al., 2020).

Another important aspect of distributed learning is managing communication costs. Indeed, communication between the nodes typically constitutes a significant portion of the time and resource consumption

[1]Following the standard terminology (Lamport et al., 1982; Su and Vaidya, 2016), we call a worker *Byzantine* if it can (maliciously or not) send incorrect information to other workers/server. Such workers are assumed to be omniscient, i.e., they have access to the vectors that other workers send, know the aggregation rule on the server, and can coordinate their actions with one another.

of the training process. Consequently, as the field of machine learning continues to leverage larger and more complex models trained on extensive datasets, the need to efficiently exchange information between nodes becomes paramount. Communication compression techniques such as quantization (Alistarh et al., 2017) and sparsification (Suresh et al., 2017; Stich et al., 2018), play a pivotal role in addressing this challenge.

While both Byzantine robustness and communication compression are individually highly significant topics, their simultaneous exploration is relatively rare in the exisitng literature. To date, only five papers have tackled both of these challenges concurrently: Bernstein et al. (2018) study signSGD with majority vote, Ghosh et al. (2020, 2021) propose methods utilising aggregation rules that select the update vectors based on their norms, and Zhu and Ling (2021); Gorbunov et al. (2023) develop variance-reduced methods. The current state-of-the-art theoretical results in this area are derived by Gorbunov et al. (2023), who propose Byz-VR-MARINA – an algorithm with provably robust aggregation (Karimireddy et al., 2021, 2022) based on the SARAH-type variance reduction (Nguyen et al., 2017), employing unbiased compression of the stochastic gradient differences (Gorbunov et al., 2021b).

The existing results have certain limitations. In particular, although Byz-VR-MARINA achieves state-of-the-art convergence rates, it requires occasional communication of uncompressed messages. Further, it has inferior theoretical guarantees for optimization error in the heterogeneous case, and tolerates less Byzantine workers in the heterogeneous over-parameterized regime compared to some other existing methods, such as Byzantine-Robust SGD with momentum (BR-SGDm) (Karimireddy et al., 2022). Moreover, existing Byzantine-robust methods only support the use of unbiased compressors. Meanwhile, it is known that employing (typically biased) contractive compressors combined with error feedback – a powerful technique introduced in the communication compression literature (Seide et al., 2014; Richtárik et al., 2021) – often yields superior empirical performance. *Our work comprehensively addresses all these limitations.*

1.1 Technical Preliminaries

In this paper, we consider a standard distributed optimization problem in which both the objective function f and the functions f_i stored on the nodes have a finite-sum structure:

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

$$\text{where}\quad f_i(x)=\frac{1}{m}\sum_{j=1}^{m}f_{i,j}(x)\quad \forall i\in\mathcal{G},$$

where $\mathcal{G}$ is the set of *good/regular/non-Byzantine* clients, $|\mathcal{G}|=G$, $f_i(x)$ corresponds to the loss of the model x on the data stored on worker i, and $f_{i,j}(x)$ is the loss on the j-th example from the local dataset of worker i. In addition to regular workers, there is a set of *bad/malicious/Byzantine* workers, denoted as $\mathcal{B}$, also participating in the training process. For notational convenience, we assume that $\mathcal{G}\cup\mathcal{B}=[n]=\{1,2,\ldots,n\}$. We refrain from making any assumptions on the behaviour of the Byzantine workers, but we do impose a constraint on their number, requiring them to constitute less than half of all clients.

Robust aggregation. One of the key ingredients of our methods is a (δ,c)-Robust Aggregator, initially introduced by Karimireddy et al. (2021, 2022). We use the generalized version from Gorbunov et al. (2023).

Definition 1.1 ((δ,c)-Robust Aggregator)**.** *Assume that $\{x_1,x_2,\ldots,x_n\}$ is such that there exists a subset $\mathcal{G}\subseteq[n]$ of size $|\mathcal{G}|=G\geq(1-\delta)n$ with $\delta<0.5$ and there exists $\sigma\geq 0$ such that $\frac{1}{G(G-1)}\sum_{i,l\in\mathcal{G}}\mathbb{E}\left[\|x_i-x_l\|^2\right]\leq\sigma^2$, where the expectation is taken w.r.t. the randomness of $\{x_i\}_{i\in\mathcal{G}}$. We say that the quantity $\widehat{x}$ is a (δ,c)-Robust Aggregator ((δ,c)-**RAgg**) and write $\widehat{x}=$ **RAgg**$(x_1,\ldots,x_n)$ for some $c>0$, if the following inequality holds:*

$$\mathbb{E}\left[\|\widehat{x}-\overline{x}\|^2\right]\leq c\delta\sigma^2, \tag{2}$$

*where $\overline{x}=\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}x_i$. If additionally $\widehat{x}$ is computed without the knowledge of σ^2, we say that $\widehat{x}$ is a (δ,c)-Agnostic Robust Aggregator ((δ,c)-**ARAgg**) and write $\widehat{x}=$ **ARAgg**$(x_1,\ldots,x_n)$.*

In essence, an aggregator is regarded as *robust* if it closely approximates the average of regular vectors. Specifically, the upper bound on the expected squared distance between the two quantities should be proportional to the pairwise variance of the non-malicious vectors, and the upper bound on the proportion of Byzantine workers. In terms of this criterion, this definition is tight (Karimireddy et al., 2021). Some examples of (δ,c)-robust aggregators are provided in Appendix B.

Communication compression. We focus on two main classes of compression operators: *unbiased* and *biased* (contractive) compressors.

Definition 1.2 (Unbiased compressor)**.** *A stochastic mapping $\mathcal{Q}:\mathbb{R}^d\to\mathbb{R}^d$ is called an* unbiased compressor/compression operator *if there exists $\omega\geq 0$ such*

that for any $x \in \mathbb{R}^d$

$$\mathbb{E}\left[\mathcal{Q}(x)\right] = x, \quad \mathbb{E}\left[\|\mathcal{Q}(x) - x\|^2\right] \leq \omega\|x\|^2. \tag{3}$$

This definition encompasses a wide range of well-known compression techniques, including Rand$K$ sparsification (Stich et al., 2018), random dithering (Goodall, 1951; Roberts, 1962) and natural compression (Horváth et al., 2019a). However, it does not cover another important class of compression operators, called contractive compressors, which are usually biased.

Definition 1.3 (Contractive compressor)**.** *A stochastic mapping* $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ *is called a* contractive compressor/compression operator *if there exists* $\alpha \in [0, 1)$ *such that for any* $x \in \mathbb{R}^d$

$$\mathbb{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq (1 - \alpha)\|x\|^2. \tag{4}$$

One of the most popular examples of contractive compressors is the Top$K$ sparsification (Alistarh et al., 2018b).

We shall denote the families of compressors satisfying Definitions 1.2 and 1.3 by $\mathbb{U}(\omega)$ and $\mathbb{B}(\alpha)$ respectively. Notably, it can easily be verified that if $\mathcal{C} \in \mathbb{U}(\omega)$, then $(\omega+1)^{-1}\mathcal{C} \in \mathbb{B}\left((\omega+1)^{-1}\right)$ (Beznosikov et al., 2020), so the family of biased compressors is wider.

Some examples of such mappings are provided in Appendix B. For a comprehensive overview of biased and unbiased compressors, we refer to the summary in (Beznosikov et al., 2020).

Assumptions. We start with formulating the standard smoothness assumption.

Assumption 1.1. *The function* $f : \mathbb{R}^d \to \mathbb{R}$ *is* L*-smooth, meaning that for any* $x, y \in \mathbb{R}^d$ *it satisfies* $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$*. In addition, we assume that* $f_* = \inf_{x\in\mathbb{R}^d} f(x) > -\infty$*.*

Our analysis also relies on a special notion of smoothness of loss functions stored on regular workers.

Assumption 1.2 (Global Hessian variance (Szlendak et al., 2021))**.** *There exists* $L_\pm \geq 0$ *such that for all* $x, y \in \mathbb{R}^d$

$$\frac{1}{G}\sum_{i\in\mathcal{G}} \|\nabla f_i(x) - \nabla f_i(y)\|^2 - \|\nabla f(x) - \nabla f(y)\|^2 \leq L_\pm^2\|x - y\|^2. \tag{5}$$

It can be verified that the above assumption is always valid for some $L_\pm \geq 0$ whenever f_i is L_i-smooth for all $i \in \mathcal{G}$. More precisely, if this is the case, then there exists $L_\pm$ satisfying the above assumption such that $L_{\text{avg}}^2 - L^2 \leq L_\pm^2 \leq L_{\text{avg}}^2$, where $L_{\text{avg}}^2 = \frac{1}{G}\sum_{i\in\mathcal{G}} L_i^2$ (Szlendak et al., 2021). Moreover, there exist problems with heterogeneous data such that (5) holds with $L_\pm = 0$, while $L_{\text{avg}} > 0$ (Szlendak et al., 2021).

The vast majority of existing works on Byzantine-robustness focus solely on the standard uniform sampling. To be able to consider a wider range of samplings, we also employ an assumption on the (expected) Lipschitzness for samplings of stochastic gradients.

Assumption 1.3 (Local Hessian variance (Gorbunov et al., 2023))**.** *There exists* $\mathcal{L}_\pm \geq 0$ *such that for all* $x, y \in \mathbb{R}^d$ *the unbiased mini-batched estimator* $\widehat{\Delta}_i(x, y)$ *of* $\Delta_i(x, y) = \nabla f_i(x) - \nabla f_i(y)$ *with batch size* b *satisfies*

$$\frac{1}{G}\sum_{i\in\mathcal{G}} \mathbb{E}\left[\|\widehat{\Delta}_i(x, y) - \Delta_i(x, y)\|^2\right] \leq \frac{\mathcal{L}_\pm^2}{b}\|x - y\|^2. \tag{6}$$

As shown in (Gorbunov et al., 2023, Appendix E.1), the above assumption is quite general, encompassing scenarios such as uniform and importance sampling of stochastic gradients.

Finally, to ensure provable Byzantine-robustness, it is necessary to introduce an assumption on heterogeneity among regular workers. Otherwise, Byzantine workers can transmit arbitrary vectors, pretending to have access to non-representative data, thus becoming undetectable.

Assumption 1.4 ((B, ζ^2)-heterogeneity)**.** *There exist* $B \geq 0$ *and* $\zeta \geq 0$ *such that for all* $x \in \mathbb{R}^d$ *the local loss functions of the good workers satisfy*

$$\frac{1}{G}\sum_{i\in\mathcal{G}} \|\nabla f_i(x) - \nabla f(x)\|^2 \leq B\|\nabla f(x)\|^2 + \zeta^2. \tag{7}$$

This assumption is the most general one of its kind in the literature on Byzantine-robustness. In particular, when $B = 0$, it reduces to the standard bounded gradient dissimilarity assumption, and when $\zeta = 0$, it implies that all stationary points of f are also stationary points of each function f_i. The latter typically occurs in over-parameterized models (Vaswani et al., 2019). Finally, it is worth mentioning that even the homogeneous case ($B = 0$, $\zeta = 0$) where all workers $i \in \mathcal{G}$ have access to the same local function $f_i = f$ is relevant, especially in collaborative learning scenarios, when data is openly shared and the aim is to speed up the training process (Diskin et al., 2021; Kijsipongse et al., 2018).

1.2 Our Contributions

Below we summarize our main contributions.

Table 1: Summary of the derived complexity bounds in the general non-convex case and a comparison with the complexity of Byz-VR-MARINA. Columns: "Rounds" = the number of communication rounds required to find x such that $\mathbb{E}\left[\|\nabla f(x)\|^2\right] \leq \varepsilon^2$; "$\varepsilon \leq$" = the lower bound for the best achievable accuracy ε; "$\delta <$" = the maximal ratio of Byzantine workers that the method can provably tolerate. Only dependencies with respect to the following variables are shown: $\omega =$ unbiased compression parameter, $n =$ number of workers, $m =$ number of local functions, $b =$ batch size, $G =$ number of regular workers, $c =$ aggregation constant, $\delta =$ percentage of Byzantine workers, $\alpha_D, \alpha_P =$ contractive compression parameters for uplink and downlink compression, respectively. The results derived in this paper are highlighted in light blue color; red color indicates terms in the complexity bound/lower bound for ε/upper bound for δ for Byz-VR-MARINA that we improve in our work.

Method	Rounds	$\varepsilon \leq$	$\delta <$
Byz-VR-MARINA[(1)] (Gorbunov et al., 2023)	$\frac{1}{\varepsilon^2}\left(1+\sqrt{\max\{(1+\omega)^2, \frac{m(1+\omega)}{b}\}}\left(\sqrt{\frac{1}{G}}+\sqrt{c\delta\max\{\omega, \frac{m}{b}\}}\right)\right)$	$\frac{c\delta\zeta^2}{p-c\delta B}$	$\frac{p}{cB}$
Byz-VR-MARINA 2.0 [(1)]	$\frac{1}{\varepsilon^2}\left(1+\sqrt{\max\{(1+\omega)^2, \frac{m(1+\omega)}{b}\}}\left(\sqrt{\frac{1}{G}}+\sqrt{c\delta}\right)\right)$	$\frac{c\delta\zeta^2}{1-c\delta B}$	$\frac{1}{(c+\sqrt{c})B}$
Byz-DASHA-PAGE [(1)]	$\frac{1}{\varepsilon^2}\left(1+\left(\omega+\frac{\sqrt{m}}{b}\right)\left(\sqrt{\frac{1}{G}}+\sqrt{c\delta}\right)\right)$	$\frac{c\delta\zeta^2}{1-c\delta B}$	$\frac{1}{(c+\sqrt{c})B}$
Byz-EF21 [(2)]	$\frac{(1+\sqrt{c\delta})}{\alpha_D\varepsilon^2}$	$\frac{(c\delta+\sqrt{c\delta})\zeta^2}{1-B(c\delta+\sqrt{c\delta})}$	$\frac{1}{c(B+B^2)}$
Byz-EF21-BC [(2)]	$\frac{(1+\sqrt{c\delta})}{\alpha_D\alpha_P\varepsilon^2}$	$\frac{(c\delta+\sqrt{c\delta})\zeta^2}{1-B(c\delta+\sqrt{c\delta})}$	$\frac{1}{c(B+B^2)}$

[(1)] These methods use unbiased compression and work with stochastic gradients. For Byz-VR-MARINA and Byz-VR-MARINA 2.0 $p = \min\{1/1+\omega, b/m\}$, for Byz-DASHA-PAGE $p = b/m$.

[(2)] These methods compute full gradients on regular workers and use (biased) contractive compression. To enable easier comparison with the methods employing unbiased compression, one can assume that the biased compressors arise from unbiased ones through scaling and the following relations hold: $1/\alpha_P = \omega_P + 1$ and $1/\alpha_D = \omega_D + 1$.

◇ **Improved complexity bounds.** We propose two new Byzantine-robust methods incorporating *unbiased* compression: Byz-VR-MARINA 2.0 and Byz-DASHA-PAGE, and prove their complexity bounds for general smooth non-convex functions under quite general assumptions about sampling and stochasticity of the gradients. The derived complexity bounds for Byz-VR-MARINA 2.0 and Byz-DASHA-PAGE improve the existing theoretical results of the current state-of-the-art Byz-VR-MARINA method, outperforming it by factors of $\sqrt{\max\{1+\omega, m/b\}}$ and $\sqrt{\max\{(1+\omega)^3, m^2(1+\omega)/b^2\}}$ in the leading term, respectively, as shown in Table 1. These are significant improvements since ω (compression parameter) and m (size of the local dataset) are usually large. Moreover, we prove that both algorithms converge linearly under the Polyak-Łojasiewicz condition (see Appendix H).

◇ **Smaller size of the neighborhood.** Under the (B, ζ^2)-heterogeneity assumption, Byz-VR-MARINA 2.0 and Byz-DASHA-PAGE converge to a smaller neighborhood of the solution than their competitors. Furthermore, when $B = 0$, our methods can achieve $\mathbb{E}[\|\nabla f(x)\|^2] = \mathcal{O}(c\delta)$, matching the lower bound by Karimireddy et al. (2022), while for Byz-VR-MARINA one can only prove $\mathbb{E}[\|\nabla f(x)\|^2] = \mathcal{O}(c\delta/p)$, which is worse by a large factor of $1/p \sim \max\{\omega, m/b\}$.

◇ **Higher tolerance to Byzantine workers.** When the (B, ζ^2)-heterogeneity assumption holds with $B > 0$, the results derived for Byz-VR-MARINA 2.0 and Byz-DASHA-PAGE guarantee convergence in the presence of $1/p$ times more Byzantine workers than in the case of Byz-VR-MARINA.

◇ **The first Byzantine-robust methods with error feedback.** Finally, we propose and analyze two new Byzantine-robust methods employing any (in general, *biased*) contractive compressors – Byz-EF21 and Byz-EF21-BC. Both are based on modern error feedback – the EF21 algorithm of Richtárik et al. (2021), and are the first provably Byzantine-robust methods utilising error feedback. The Byz-EF21-BC algorithm, in addition to workers-to-server compression, also compresses messages sent from the server to workers, hence being the first provably Byzantine-robust algorithm using bidirectional compression.

1.3 Related Work on Byzantine Robustness

The first approaches to designing Byzantine-robust distributed optimization methods[2] primarily concentrate on the aggregation aspect, employing conventional Parallel-SGD as the algorithm's foundation (Blanchard et al., 2017a; Chen et al., 2017; Yin et al., 2018; Damaskinos et al., 2019; Guerraoui et al., 2018; Pillutla et al., 2022). Nonetheless, it has become evident that these classical approaches are susceptible to specialized attacks (Baruch et al., 2019; Xie et al., 2020). To address this issue, Karimireddy et al. (2021) develop a formal definition of robust aggregation and propose a provably Byzantine-robust algorithm based on the aggregation of client momentums. Karimireddy

[2] We defer the discussion of related work on communication compression to Appendix A.

et al. (2022) further extend these results to the heterogeneous setup. An alternative formalism of robust aggregation, along with new examples of robust aggregation rules and Byzantine-robust methods, are proposed and analyzed by Allouah et al. (2023). The application of variance reduction to achieve Byzantine robustness is first explored by Wu et al. (2020), who use SAGA-type variance reduction (Defazio et al., 2014) on regular workers. Subsequently, Zhu and Ling (2021) enhance this approach by incorporating unbiased communication compression. Gorbunov et al. (2023) improve the results from Wu et al. (2020); Zhu and Ling (2021) and derive the current theoretical state-of-the-art convergence results. Numerous other approaches have also been proposed, including the banning of Byzantine workers (Alistarh et al., 2018a; Allen-Zhu et al., 2021), random checks of computations (Gorbunov et al., 2021a), computation redundancy (Chen et al., 2018; Rajput et al., 2019), and reputation scores (Rodríguez-Barroso et al., 2020; Regatti et al., 2020; Xu and Lyu, 2020). We refer to Lyu et al. (2020) for a comprehensive survey.

2 METHODS WITH UNBIASED COMPRESSION

In this section, we introduce our main results on methods employing unbiased compression.

2.1 Warm-up: Byz-VR-MARINA 2.0

We begin by presenting Byz-VR-MARINA 2.0 (Algorithm 1) – a modification of Byz-VR-MARINA by Gorbunov et al. (2023) that uses local vectors g_i^t instead of g^t in the update of g_i^{t+1}. In other words, line 12 of the original Byz-VR-MARINA method is $g_i^{t+1} = g^t + m_i^{t+1}$. The key idea behind both versions remains the same: to adapt VR-MARINA (Gorbunov et al., 2021b) by replacing the standard averaging of gradient estimators with a (δ, c)-agnostic robust aggregator. It is worth mentioning that even without this modification and with no Byzantine workers, the gradient estimator in VR-MARINA is *conditionally biased*, i.e., $\mathbb{E}[g_i^{t+1} \mid x^{t+1}, x^t] \neq \nabla f_i(x^{t+1})$. With this insight in mind, and noting that in the homogeneous scenario, the variance of vectors received from regular workers converges to zero, robust aggregation naturally integrates with the algorithm, enabling provable tolerance to Byzantine workers.

While the algorithmic difference between Byz-VR-MARINA and Byz-VR-MARINA 2.0 is almost negligible, it provides significant theoretical improvements, as demonstrated in Table 1. This advancement has an intuitive explanation: g^t contains the information received from Byzantine workers in the previous steps,

Algorithm 1 Byz-VR-MARINA 2.0

Input: starting point $x_0 \in \mathbb{R}^d$, stepsize $\gamma > 0$, probability $p \in (0,1]$, number of iterations $T \geq 1$, a collection of unbiased compressors $\{\mathcal{Q}_i\}_{i \in \mathcal{G}}$

for $t = 0, 1, \ldots, T-1$ **do**
  Let $c^{t+1} = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1-p \end{cases}$
  Broadcast g^t to all nodes
  for $i \in \mathcal{G}$ in parallel **do**
    $x^{t+1} = x^t - \gamma g^t$
    if $c^{t+1} = 1$ **then**
      $g_i^{t+1} = \nabla f_i(x^{t+1})$
      Send $\nabla f_i(x^{t+1})$ to the server.
    else
      $m_i^{t+1} = \mathcal{Q}_i(\widehat{\Delta}_i(x^{t+1}, x^t))$
      $g_i^{t+1} = g_i^t + m_i^{t+1}$
      Send m_i^{t+1} to the server
    end if
  end for
  $g^{t+1} = \text{ARAgg}\left(g_1^{t+1}; \ldots; g_n^{t+1}\right)$
end for

so the vectors $\{g_i^{t+1}\}_{i \in \mathcal{G}}$ in Byz-VR-MARINA are more influenced by malicious messages than the vectors $\{g_i^{t+1}\}_{i \in \mathcal{G}}$ in Byz-VR-MARINA 2.0. Moreover, it can also be shown that in Byz-VR-MARINA 2.0 the local gradient estimators are unbiased, so that $\mathbb{E}[g_i^{t+1}] = \mathbb{E}[\nabla f_i(x^{t+1})]$ for all $i \in \mathcal{G}$. This is not the case in Byz-VR-MARINA. Despite the apparent similarity of our algorithm to Byz-VR-MARINA, the underlying intuition significantly differ. We refer to Appendix D for further details.

The following theorem formalizes the main convergence result for Byz-VR-MARINA 2.0.

Theorem 2.1. *Let Assumptions 1.1, 1.2, 1.3 and 1.4 hold. Assume that* $0 < \gamma \leq (L + \sqrt{\eta})^{-1}$, $\delta < \left((8c + 4\sqrt{c})B\right)^{-1}$ *and initialize* $g_i^0 = \nabla f_i(x^0)$ *for all* $i \in \mathcal{G}$, *where* $\eta = \frac{1-p}{p}\left(\omega\left(\frac{\mathcal{L}_\pm^2}{b} + L_\pm^2 + L^2\right) + \frac{\mathcal{L}_\pm^2}{b}\right)\left(\sqrt{\frac{1}{G}} + \sqrt{8c\delta}\right)^2$. *Then for all* $T \geq 0$ *the iterates produced by* Byz-VR-MARINA 2.0 *satisfy*

$$\mathbb{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right] \leq \frac{1}{A}\left(\frac{2\delta^0}{\gamma T} + \left(8c\delta + \sqrt{\frac{8c\delta}{G}}\right)\zeta^2\right),$$

where $\delta^0 = f(x^0) - f^*$, $A = 1 - \left(8c\delta + \sqrt{8c\delta/G}\right)B$ *and* $\widehat{x}^T$ *is chosen uniformly at random from* $x^0, x^1, \ldots, x^{T-1}$.

The above upper bound consists of two terms: the first one decreases to zero at a rate $\mathcal{O}(1/T)$, which is

Algorithm 2 Byz-DASHA-PAGE

Input: starting point $x^0 \in \mathbb{R}^d$, stepsize $\gamma > 0$, momentum $a \in (0,1]$, probability $p \in (0,1]$, number of iterations $T \geq 1$, a collection of unbiased compressors $\{\mathcal{Q}_i\}_{i\in\mathcal{G}}$
for $t = 0, 1, \ldots, T-1$ **do**
  Let $c^{t+1} = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1-p \end{cases}$
  Broadcast g^t to all nodes
  for $i \in \mathcal{G}$ in parallel **do**
    $x^{t+1} = x^t - \gamma g^t$
    if $c^{t+1} = 1$ **then**
      $h_i^{t+1} = \nabla f_i(x^{t+1})$
    else
      $h_i^{t+1} = h_i^t + \widehat{\Delta}_i(x^{t+1}, x^t)$
    end if
    $m_i^{t+1} = \mathcal{Q}_i(h_i^{t+1} - h_i^t - a\,(g_i^t - h_i^t))$
    $g_i^{t+1} = g_i^t + m_i^{t+1}$
    Send m_i^{t+1} to the server
  end for
  $g^{t+1} = \mathrm{ARAgg}(g_1^{t+1}; \ldots; g_n^{t+1})$
end for

optimal for approximating first-order stationary points in the setup we consider (Fang et al., 2018; Arjevani et al., 2023), and the second one, corresponding to the size of the neighbourhood of the solution to which the method converges, is constant. This neighbourhood disappears when either $\zeta = 0$ or $\delta = 0$ (no malicious workers). Furthermore, when $B = 0$, the second term is $\mathcal{O}(\delta\zeta^2)$ (due to the fact that $1/G = \mathcal{O}(\delta)$ when $\delta > 0$), matching the lower bound form Karimireddy et al. (2022) up to a numerical factor. Meanwhile, the corresponding term in the results derived for Byz-VR-MARINA is $\mathcal{O}(\delta\zeta^2/p)$, which is usually much larger than $\mathcal{O}(\delta\zeta^2)$ since p is typically small. When $B > 0$, the results for Byz-VR-MARINA 2.0 are valid when $\delta = \mathcal{O}(1/B)$, while the existing guarantees for Byz-VR-MARINA require $\delta = \mathcal{O}(p/B)$, which is significantly smaller. Finally, as shown in Table 1, the rate of convergence of Byz-VR-MARINA 2.0 is better than that of Byz-VR-MARINA by a potentially large factor of $\sqrt{\max\{\omega, m/b\}}$.

2.2 Byz-DASHA-PAGE

Although Byz-VR-MARINA 2.0 enjoys notable theoretical improvements, it inherits some limitations of VR-MARINA. The first one is of purely algorithmic nature: with probability p, regular workers send uncompressed vectors. These synchronization steps are necessary for the algorithm to converge, as they correct the error coming from compression and stochasticity in the gradients. However, their large communication cost can render the algorithm impractical. To make the computation and communication cost of one round equal (on average, up to a constant factor) to the cost per round when workers compute stochastic gradients and send compressed vectors, the probability p should be approximately $\min\{(1+\omega)^{-1}, b/m\}$. If this is the case, a factor of $\sqrt{1+\omega}\sqrt{\max\{1+\omega, m/b\}}$ appears in the complexity bounds of (Byz-)VR-MARINA (2.0), and the effects of stochasticity and communication compression are coupled.

To address the issues arising in VR-MARINA, Tyurin and Richtárik (2023b) propose DASHA-PAGE, which uses momentum variance reduction mechanisms (Cutkosky and Orabona, 2019; Tran-Dinh et al., 2022; Liu et al., 2020) to handle the noise resulting from communication compression (and a *separate* Geom-SARAH/PAGE-type variance reduction technique to manage the stochasticity in the gradients). The derived complexity of DASHA-PAGE matches that of VR-MARINA, but with a better dependence on ω, m, b: the factor $\sqrt{1+\omega}\sqrt{\max\{1+\omega, m/b\}}$ appearing in the complexity bounds of VR-MARINA is replaced with $\omega + \frac{\sqrt{m}}{b}$, which is always not greater than $\sqrt{1+\omega}\sqrt{\max\{1+\omega, m/b\}}$ and strictly smaller than $\sqrt{1+\omega}\sqrt{\max\{1+\omega, m/b\}}$ when $0 < \omega < m/b$.

Motivated by these developments, we introduce a Byzantine-robust variant of DASHA-PAGE called Byz-DASHA-PAGE (Algorithm 2), aiming to enhance the convergence rates achieved by Byz-VR-MARINA 2.0. We find that (δ, c)-robust aggregation integrates seamlessly with DASHA-PAGE, leading to the following result.

Theorem 2.2. *Let Assumptions 1.1, 1.2, 1.3 and 1.4 hold. Assume that* $0 < \gamma \leq (L + \sqrt{\eta})^{-1}$, $\delta < \left((8c + 4\sqrt{c})B\right)^{-1}$ *and initialize* $g_i^0 = \nabla f_i(x^0)$ *for all* $i \in \mathcal{G}$, *where* $\eta = \left(8\omega(2\omega+1)\left(L_\pm^2 + L^2\right) + \frac{1-p}{b}\left(12\omega\,(2\omega+1) + \frac{2}{p}\right)\mathcal{L}_\pm^2\right) \times \left(\sqrt{\frac{1}{G}} + \sqrt{8c\delta}\right)^2$. *Then for all* $T \geq 0$ *the iterates produced by* Byz-DASHA-PAGE *satisfy*

$$\mathbb{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right] \leq \frac{1}{A}\left(\frac{2\delta^0}{\gamma T} + \left(8c\delta + \sqrt{\frac{8c\delta}{G}}\right)\zeta^2\right),$$

where $\delta^0 = f(x^0) - f^*$, $A = 1 - \left(8c\delta + \sqrt{8c\delta/G}\right)B$ *and* $\widehat{x}^T$ *is chosen uniformly at random from* $x^0, x^1, \ldots, x^{T-1}$.

In terms of the size of the neighborhood and maximum number of Byzantine workers, the above result aligns closely with what we obtain for Byz-VR-MARINA 2.0, thereby being superior to the existing guarantees for Byz-VR-MARINA. However, the upper bound

Algorithm 3 Byz-EF21

Input: starting point $x^0 \in \mathbb{R}^d$, stepsize $\gamma > 0$, number of iterations $T \geq 1$, a collection of biased compressors $\{\mathcal{C}_i\}_{i\in\mathcal{G}}$
for $t = 0, 1, \ldots, T-1$ **do**
  $x^{t+1} = x^t - \gamma g^t$
  Broadcast x^{t+1} to all workers
  for $i \in \mathcal{G}$ **in parallel do**
    $c_i^t = \mathcal{C}_i(\nabla f_i(x^{t+1}) - g_i^t)$
    $g_i^{t+1} = g_i^t + c_i^t$
    Send message c_i^t to the server
  end for
  $g^{t+1} = \texttt{ARAgg}(g_1^{t+1}, \ldots, g_n^{t+1})$
end for

Algorithm 4 Byz-EF21-BC

Input: starting point $x^0 \in \mathbb{R}^d$, stepsize $\gamma > 0$, number of iterations $T \geq 1$, a collection of biased compressors $\{\mathcal{C}_i^D\}_{i\in\mathcal{G}}$, $\mathcal{C}^P$
for $t = 0, 1, \ldots, T-1$ **do**
  $x^{t+1} = x^t - \gamma g^t$
  $s^{t+1} = \mathcal{C}^P\left(x^{t+1} - w^t\right)$
  $w^{t+1} = w^t + s^{t+1}$
  Broadcast s^{t+1} to all workers
  for $i \in \mathcal{G}$ **in parallel do**
    $w^{t+1} = w^t + s^{t+1}$
    $c_i^t = \mathcal{C}_i^D(\nabla f_i(w^{t+1}) - g_i^t)$
    $g_i^{t+1} = g_i^t + c_i^t$
    Send message c_i^t to the server
  end for
  $g^{t+1} = \text{ARAgg}(g_1^{t+1}, \ldots, g_n^{t+1})$
end for

on the stepsize of Byz-DASHA-PAGE has a better joint dependence on ω and p compared to that in Byz-VR-MARINA 2.0. The difference can be seen in the expression for η: in Theorem 2.2, $1/p$ and ω are decoupled, whereas η from Theorem 2.1 has a term proportional to $(1+\omega)/p$, ultimately leading to the factor of $\sqrt{1+\omega}\sqrt{\max\{1+\omega, m/b\}}$ in the complexity bounds of Byz-VR-MARINA 2.0. A comparison of complexity results is given in Table 1.

3 METHODS WITH BIASED COMPRESSION AND ERROR FEEDBACK

In this section, we transition from a discussion of methods using unbiased compressors to algorithms utilizing (typically biased) contractive compressors, which typically have better empirical performance than the unbiased alternatives (Seide et al., 2014). Such compressors are usually employed together with error feedback mechanisms, since their naive use in distributed Gradient Descent can lead to divergence of the algorithm (Beznosikov et al., 2020). In this work, we focus on the modern EF21 error feedback mechanism proposed by Richtárik et al. (2021), as it offers better convergence guarantees compared to standard error feedback. The method is based on the idea of each worker compressing the difference between the current gradient and its estimate g_i^t and using this compressed message to update the local gradient estimate in the next round. Since both $\nabla f_i(x^t)$ and g_i^t converge to $\nabla f_i(x^*)$ (where x^* is a stationary point to which the method converges), $\nabla f_i(x^t) - g_i^t$ tends to zero. Given that the compressor is contractive, it must be the case that the inaccuracy due to the compression of this difference also converges to zero. Importantly, from an algorithmic point of view, EF21 resembles MARINA, which is known to work well with robust aggregation.

Motivated by the above considerations, we propose two new Byzantine-robust methods – Byz-EF21 and Byz-EF21-BC (Algorithms 3 and 4). Byz-EF21 is a modification of the EF21 mechanism employing (δ, c)-robust aggregation to ensure Byzantine-robustness. Byz-EF21-BC further enhances the method by adding bidirectional compression: following Gruntkowska et al. (2023), we additionally apply the EF21 mechanism on the server's side to compress messages broadcast to workers. Note that when $\mathcal{C}^P$ is the identity operator, the latter algorithm reduces to Byz-EF21. The intuition behind both algorithms mirrors that of Byz-VR-MARINA (2.0): the variance of the estimates $\{g_i^t\}_{i\in\mathcal{G}}$ approaches $\mathcal{O}(\zeta^2)$, leaving Byzantine workers progressively less room to "hide in the noise".

The following theorem presents the main result for Byz-EF21 and Byz-EF21-BC in a unified manner.

Theorem 3.1. *Let Assumptions 1.1, 1.2, and 1.4 hold. Assume that* $\mathcal{C}_i^D \in \mathbb{B}(\alpha_D)$, $\mathcal{C}^P \in \mathbb{B}(\alpha_P)$, $0 < \gamma \leq (L + \sqrt{\eta})^{-1}$ *and* $\delta < (8c(\sqrt{B}+B)^2)^{-1}$, *where* $\eta = \frac{32}{\alpha_D^2}\left(1 + \frac{5}{\alpha_P^2}\right)\left(1 + \sqrt{8c\delta}\right)^2\left(L_\pm^2 + L^2\right)$. *Initialize* $w^0 = x^0$, *and* $g_i^0 = \nabla f_i(x^0)$ *for all* $i \in \mathcal{G}$. *Then for all* $T \geq 0$, *the iterates produced by* Byz-EF21/Byz-EF21-BC *satisfy*

$$\mathbb{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right] \leq \frac{1}{A}\left(\frac{2\delta^0}{\gamma T} + \left(8c\delta + \sqrt{8c\delta}\right)\zeta^2\right),$$

where $\delta^0 = f(x^0) - f^*$, $A = 1 - B\left(8c\delta + \sqrt{8c\delta}\right)$ *and* $\widehat{x}^T$ *is chosen uniformly at random from* $x^0, x^1, \ldots, x^{T-1}$.

Similar to the bounds we derived for Byz-VR-MARINA

2.0 and Byz-DASHA-PAGE, the above upper bound for Byz-EF21/Byz-EF21-BC has two terms: the first decreases at a rate $\mathcal{O}(1/T)$, and the second remains constant. The neighbourhood again vanishes when either $\zeta = 0$ or $\delta = 0$. In the scenario where $B = 0$ and $\sqrt{c\delta} = \mathcal{O}(c\delta)$ (which occurs when there are many Byzantine workers), the second term is $\mathcal{O}(\delta\zeta^2)$, matching the lower bound from Karimireddy et al. (2022). Furthermore, when $\delta = 0$ (no Byzantines), the rate aligns with the result for EF21-BC (Fatkhullin et al., 2021), and when additionally $\alpha_P = 1$ (no downlink compression), the rate matches the one of EF21 (Richtárik et al., 2021). Finally, when $B > 0$, the above result holds whenever $\delta < 1/8c(\sqrt{B}+B)^2$. While this bound is worse than the one we derive for Byz-VR-MARINA 2.0 and Byz-DASHA-PAGE, it is better than that of Byz-VR-MARINA when $B < 1/p$ (which occurs, for example, when ω is large and B is small).

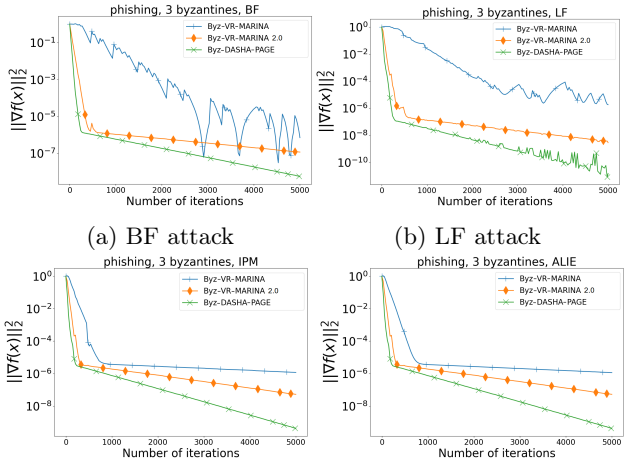
4 NUMERICAL EXPERIMENTS

We conduct an empirical comparison[3] of Byz-VR-MARINA, Byz-VR-MARINA 2.0 and Byz-DASHA-PAGE in both homogeneous and heterogeneous settings. More details and additional experiments, including those on error feedback methods, are provided in Appendix I.

We solve a binary logistic regression problem with non-convex regularizer, using the `phishing` dataset from `LibSVM` (Chang and Lin, 2011). The data is divided among $n = 16$ workers, out of which 3 are Byzantine. As the aggregation rule, we use the Coordinate-wise Median (CM) aggregator (Yin et al., 2018) with bucketing (Karimireddy et al., 2022) (see Appendix B). We consider four different attacks performed by the Byzantine clients: *Bit Flipping* (BF): change the sign of the update, *Label Flipping* (LF): change the labels, i.e., $y_{i,j} \mapsto -y_{i,j}$, *Inner Product Manipulation* (IPM): send $-\frac{z}{G}\sum_{i\in\mathcal{G}}\nabla f_i(x)$, and *A Little Is Enough* (ALIE): estimate the mean $\mu_{\mathcal{G}}$ and standard deviation $\sigma_{\mathcal{G}}$ of the regular updates and send $\mu_{\mathcal{G}} - z\sigma_{\mathcal{G}}$, where z is a constant controlling the strength of the attacks.

Non-convex homogeneous setting. In the first set of experiments (Figure 1), we showcase the efficiency of our methods in the homogeneous setting ($B = \zeta = 0$ in Assumption 1.4). Regardless of the attack type, Byz-VR-MARINA 2.0 and Byz-DASHA-PAGE significantly outperform the Byz-VR-MARINA baseline, both in terms of convergence speed and achieved accuracy, with Byz-DASHA-PAGE taking the lead.

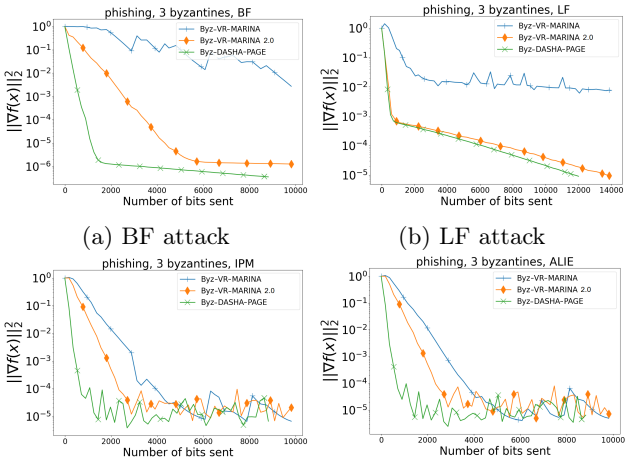
[3]Our codes are available online: https://github.com/Nikosimus/CC-for-BR-Learning.



(a) BF attack (b) LF attack

(c) IPM attack (d) ALIE attack

Figure 1: Convergence in terms of the number of iterations in the homogeneous non-convex setting.



(a) BF attack (b) LF attack

(c) IPM attack (d) ALIE attack

Figure 2: Convergence in terms of the number of bits sent in the heterogeneous non-convex setting.

Non-convex heterogeneous setting. As in the homogeneous scenario, both Byz-VR-MARINA 2.0 and Byz-DASHA-PAGE converge faster than Byz-VR-MARINA (Figure 2). The empirical superiority of our methods is especially apparent in the case of BF and LF attacks. Importantly, as suggested by theory, in the heterogeneous setting ($B, \zeta > 0$ in Assumption 1.4), the algorithms converge only to a certain neighborhood of the solution. The radius of this neighborhood is larger in the case of Byz-VR-MARINA compared to our proposed alternatives. A noteworthy observation is the inherent stability of Byz-VR-MARINA 2.0 and Byz-DASHA-PAGE, as they consistently exhibit less fluctuations compared to Byz-VR-MARINA.

Acknowledgements

The work of A. Rammal was performed during a research internship at KAUST led by P. Richtárik. A. Rammal is a student at École Polytechnique, France. The work of N. Fedin was 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.

References

Alistarh, D., Allen-Zhu, Z., and Li, J. (2018a). Byzantine stochastic gradient descent. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 4618–4628.

Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. (2017). QSGD: Communication-efficient sgd via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, 30.

Alistarh, D., Hoefler, T., Johansson, M., Konstantinov, N., Khirirat, S., and Renggli, C. (2018b). The convergence of sparsified gradient methods. *Advances in Neural Information Processing Systems*, 31.

Allen-Zhu, Z., Ebrahimian, F., Li, J., and Alistarh, D. (2021). Byzantine-resilient non-convex stochastic gradient descent. In *International Conference on Learning Representations*.

Allouah, Y., Farhadkhani, S., Guerraoui, R., Gupta, N., Pinot, R., and Stephan, J. (2023). Fixing by mixing: A recipe for optimal byzantine ml under heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pages 1232–1300. PMLR.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. (2023). Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2):165–214.

Baruch, G., Baruch, M., and Goldberg, Y. (2019). A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems*, 32.

Basu, D., Data, D., Karakus, C., and Diggavi, S. (2019). Qsparse-local-sgd: Distributed sgd with quantization, sparsification and local computations. *Advances in Neural Information Processing Systems*, 32.

Bernstein, J., Zhao, J., Azizzadenesheli, K., and Anandkumar, A. (2018). signsgd with majority vote is communication efficient and fault tolerant. *arXiv preprint arXiv:1810.05291*.

Beznosikov, A., Gorbunov, E., Berard, H., and Loizou, N. (2022). Stochastic gradient descent-ascent: Unified theory and new efficient methods. *arXiv preprint arXiv:2202.07262*.

Beznosikov, A., Horváth, S., Richtárik, P., and Safaryan, M. (2020). On biased compression for distributed learning. *arXiv preprint arXiv:2002.12410*.

Beznosikov, A., Richtárik, P., Diskin, M., Ryabinin, M., and Gasnikov, A. (2021). Distributed methods with compressed communication for solving variational inequalities, with theoretical guarantees. *arXiv preprint arXiv:2110.03313*.

Blanchard, P., El Mhamdi, E. M., Guerraoui, R., and Stainer, J. (2017a). Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in Neural Information Processing Systems*, 30.

Blanchard, P., El Mhamdi, E. M., Guerraoui, R., and Stainer, J. (2017b). Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27.

Chen, L., Wang, H., Charles, Z., and Papailiopoulos, D. (2018). Draco: Byzantine-resilient distributed training via redundant gradients. In *International Conference on Machine Learning*, pages 903–912. PMLR.

Chen, Y., Su, L., and Xu, J. (2017). Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):1–25.

Cutkosky, A. and Orabona, F. (2019). Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32.

Damaskinos, G., El-Mhamdi, E.-M., Guerraoui, R., Guirguis, A., and Rouault, S. (2019). Aggregathor: Byzantine machine learning via robust gradient aggregation. *Proceedings of Machine Learning and Systems*, 1:81–106.

Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27.

Diskin, M., Bukhtiyarov, A., Ryabinin, M., Saulnier, L., Sinitsin, A., Popov, D., Pyrkin, D. V., Kashirin, M., Borzunov, A., Villanova del Moral, A., et al. (2021). Distributed deep learning in open collaborations. *Advances in Neural Information Processing Systems*, 34:7879–7897.

Faghri, F., Tabrizian, I., Markov, I., Alistarh, D., Roy, D. M., and Ramezani-Kebrya, A. (2020). Adaptive gradient quantization for data-parallel sgd. *Advances in neural information processing systems*, 33:3174–3185.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. (2018). Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31.

Fang, M., Cao, X., Jia, J., and Gong, N. (2020). Local model poisoning attacks to {Byzantine-Robust} federated learning. In *29th USENIX security symposium (USENIX Security 20)*, pages 1605–1622.

Fatkhullin, I., Sokolov, I., Gorbunov, E., Li, Z., and Richtárik, P. (2021). EF21 with bells & whistles: Practical algorithmic extensions of modern error feedback. *arXiv preprint arXiv:2110.03294*.

Ghosh, A., Maity, R. K., Kadhe, S., Mazumdar, A., and Ramchandran, K. (2021). Communication-efficient and byzantine-robust distributed learning with error feedback. *IEEE Journal on Selected Areas in Information Theory*, 2(3):942–953.

Ghosh, A., Maity, R. K., and Mazumdar, A. (2020). Distributed newton can communicate less and resist byzantine workers. *Advances in Neural Information Processing Systems*, 33:18028–18038.

Goodall, W. (1951). Television by pulse code modulation. *Bell System Technical Journal*, 30(1):33–49.

Gorbunov, E., Borzunov, A., Diskin, M., and Ryabinin, M. (2021a). Secure distributed training at scale. *arXiv preprint arXiv:2106.11257*.

Gorbunov, E., Burlachenko, K. P., Li, Z., and Richtárik, P. (2021b). MARINA: Faster non-convex distributed learning with compression. In *International Conference on Machine Learning*, pages 3788–3798. PMLR.

Gorbunov, E., Horváth, S., Richtárik, P., and Gidel, G. (2023). Variance reduction is an antidote to byzantines: Better rates, weaker assumptions and communication compression as a cherry on the top.

Gorbunov, E., Kovalev, D., Makarenko, D., and Richtárik, P. (2020). Linearly converging error compensated sgd. *Advances in Neural Information Processing Systems*, 33:20889–20900.

Gruntkowska, K., Tyurin, A., and Richtárik, P. (2023). Ef21-p and friends: Improved theoretical communication complexity for distributed optimization with bidirectional compression. In *International Conference on Machine Learning*, pages 11761–11807. PMLR.

Guerraoui, R., Rouault, S., et al. (2018). The hidden vulnerability of distributed learning in byzantium. In *International Conference on Machine Learning*, pages 3521–3530. PMLR.

Haddadpour, F., Kamani, M. M., Mokhtari, A., and Mahdavi, M. (2021). Federated learning with compression: Unified analysis and sharp guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 2350–2358. PMLR.

Horváth, S., Ho, C.-Y., Horvath, L., Sahu, A. N., Canini, M., and Richtárik, P. (2019a). Natural compression for distributed deep learning. *arXiv preprint arXiv:1905.10988*.

Horváth, S., Kovalev, D., Mishchenko, K., Stich, S., and Richtárik, P. (2019b). Stochastic distributed learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*.

Islamov, R., Qian, X., and Richtárik, P. (2021). Distributed second order methods with fast rates and compressed communication. In *International Conference on Machine Learning*, pages 4617–4628. PMLR.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210.

Karimireddy, S. P., He, L., and Jaggi, M. (2021). Learning from history for byzantine robust optimization. In *International Conference on Machine Learning*, pages 5311–5319. PMLR.

Karimireddy, S. P., He, L., and Jaggi, M. (2022). Byzantine-robust learning on heterogeneous datasets via bucketing. *International Conference on Learning Representations*.

Karimireddy, S. P., Rebjock, Q., Stich, S., and Jaggi, M. (2019). Error feedback fixes signsgd and other gradient compression schemes. In *International Conference on Machine Learning*, pages 3252–3261. PMLR.

Khirirat, S., Feyzmahdavian, H. R., and Johansson, M. (2018). Distributed learning with compressed gradients. *arXiv preprint arXiv:1806.06573*.

Kijsipongse, E., Piyatumrong, A., et al. (2018). A hybrid gpu cluster and volunteer computing platform

for scalable deep learning. *The Journal of Supercomputing*, 74(7):3236–3263.

Koloskova, A., Stich, S., and Jaggi, M. (2019). Decentralized stochastic optimization and gossip algorithms with compressed communication. In *International Conference on Machine Learning*, pages 3478–3487. PMLR.

Konečný, J., McMahan, H. B., Yu, F., Richtárik, P., Suresh, A. T., and Bacon, D. (2016). Federated learning: strategies for improving communication efficiency. In *NIPS Private Multi-Party Machine Learning Workshop*.

Kovalev, D., Koloskova, A., Jaggi, M., Richtarik, P., and Stich, S. (2021). A linearly convergent algorithm for decentralized optimization: Sending less bits for free! In *International Conference on Artificial Intelligence and Statistics*, pages 4087–4095. PMLR.

Lamport, L., Shostak, R., and Pease, M. (1982). The byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3):382–401.

Li, C. (2020). Demystifying gpt-3 language model: A technical overview.

Li, S., Ngai, E. C.-H., and Voigt, T. (2023). An experimental study of Byzantine-robust aggregation schemes in federated learning. *IEEE Transactions on Big Data*.

Li, Z., Bao, H., Zhang, X., and Richtárik, P. (2021). PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, pages 6286–6295. PMLR.

Li, Z., Kovalev, D., Qian, X., and Richtarik, P. (2020). Acceleration for compressed gradient descent in distributed and federated optimization. In *International Conference on Machine Learning*, pages 5895–5904. PMLR.

Li, Z. and Richtárik, P. (2021). Canita: Faster rates for distributed convex optimization with communication compression. *Advances in Neural Information Processing Systems*, 34.

Liu, D., Nguyen, L. M., and Tran-Dinh, Q. (2020). An optimal hybrid variance-reduced algorithm for stochastic composite nonconvex optimization. *arXiv preprint arXiv:2008.09055*.

Lyu, L., Yu, H., Ma, X., Sun, L., Zhao, J., Yang, Q., and Yu, P. S. (2020). Privacy and robustness in federated learning: Attacks and defenses. *arXiv preprint arXiv:2012.06337*.

Mishchenko, K., Gorbunov, E., Takáč, M., and Richtárik, P. (2019). Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. (2017). Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR.

OpenAI (2023). GPT-4 technical report.

Özfatura, K., Özfatura, E., Küpçü, A., and Gündüz, D. (2023). Byzantines can also learn from history: Fall of centered clipping in federated learning. *IEEE Transactions on Information Forensics and Security*.

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

Philippenko, C. and Dieuleveut, A. (2021). Preserved central model for faster bidirectional compression in distributed settings. *Advances in Neural Information Processing Systems*, 34.

Pillutla, K., Kakade, S. M., and Harchaoui, Z. (2022). Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154.

Qian, X., Richtárik, P., and Zhang, T. (2021). Error compensated distributed sgd can be accelerated. *Advances in Neural Information Processing Systems*, 34.

Rajput, S., Wang, H., Charles, Z., and Papailiopoulos, D. (2019). Detox: A redundancy-based framework for faster and more robust gradient aggregation. *Advances in Neural Information Processing Systems*, 32.

Regatti, J., Chen, H., and Gupta, A. (2020). ByGARS: Byzantine SGD with arbitrary number of attackers. *arXiv preprint arXiv:2006.13421*.

Richtárik, P., Sokolov, I., and Fatkhullin, I. (2021). EF21: A new, simpler, theoretically better, and practically faster error feedback. *Advances in Neural Information Processing Systems*, 34.

Roberts, L. (1962). Picture coding using pseudo-random noise. *IRE Transactions on Information Theory*, 8(2):145–154.

Rodríguez-Barroso, N., Martínez-Cámara, E., Luzón, M., Seco, G. G., Veganzones, M. Á., and Herrera, F. (2020). Dynamic federated learning model for identifying adversarial clients. *arXiv preprint arXiv:2007.15030*.

Sadiev, A., Malinovsky, G., Gorbunov, E., Sokolov, I., Khaled, A., Burlachenko, K., and Richtárik, P.

(2022). Federated optimization algorithms with random reshuffling and gradient compression. *arXiv preprint arXiv:2206.07021*.

Safaryan, M., Islamov, R., Qian, X., and Richtárik, P. (2021). Fednl: Making newton-type methods applicable to federated learning. *arXiv preprint arXiv:2106.02969*.

Sahu, A., Dutta, A., M Abdelmoniem, A., Banerjee, T., Canini, M., and Kalnis, P. (2021). Rethinking gradient sparsification as total error minimization. *Advances in Neural Information Processing Systems*, 34.

Seide, F., Fu, H., Droppo, J., Li, G., and Yu, D. (2014). 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*. Citeseer.

Stich, S. U., Cordonnier, J.-B., and Jaggi, M. (2018). Sparsified sgd with memory. *Advances in Neural Information Processing Systems*, 31.

Su, L. and Vaidya, N. H. (2016). Fault-tolerant multi-agent optimization: optimal iterative distributed algorithms. In *Proceedings of the 2016 ACM symposium on principles of distributed computing*, pages 425–434.

Suresh, A. T., Felix, X. Y., Kumar, S., and McMahan, H. B. (2017). Distributed mean estimation with limited communication. In *International Conference on Machine Learning*, pages 3329–3337. PMLR.

Szlendak, R., Tyurin, A., and Richtárik, P. (2021). Permutation compressors for provably faster distributed nonconvex optimization. *arXiv preprint arXiv:2110.03300*.

Tang, H., Yu, C., Lian, X., Zhang, T., and Liu, J. (2019). DoubleSqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. In *International Conference on Machine Learning*, pages 6155–6165.

Tao, Y., Cui, S., Xu, W., Yin, H., Yu, D., Liang, W., and Cheng, X. (2023). Byzantine-resilient federated learning at edge. *IEEE Transactions on Computers*.

Tran-Dinh, Q., Pham, N. H., Phan, D. T., and Nguyen, L. M. (2022). A hybrid stochastic optimization framework for composite nonconvex optimization. *Mathematical Programming*, 191(2):1005–1071.

Tyurin, A. and Richtárik, P. (2023a). 2direction: Theoretically faster distributed training with bidirectional communication compression. *arXiv preprint arXiv:2305.12379*.

Tyurin, A. and Richtárik, P. (2023b). DASHA: Distributed nonconvex optimization with communication compression, optimal oracle complexity, and no client synchronization. *International Conference on Learning Representations*.

Vaswani, S., Bach, F., and Schmidt, M. (2019). Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pages 1195–1204. PMLR.

Vogels, T., Karimireddy, S. P., and Jaggi, M. (2019). Powersgd: Practical low-rank gradient compression for distributed optimization. *Advances in Neural Information Processing Systems*, 32.

Weiszfeld, E. (1937). Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematical Journal, First Series*, 43:355–386.

Wen, W., Xu, C., Yan, F., Wu, C., Wang, Y., Chen, Y., and Li, H. (2017). Terngrad: Ternary gradients to reduce communication in distributed deep learning. *Advances in neural information processing systems*, 30.

Wu, Z., Ling, Q., Chen, T., and Giannakis, G. B. (2020). Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks. *IEEE Transactions on Signal Processing*, 68:4583–4596.

Xie, C., Koyejo, O., and Gupta, I. (2020). Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation. In *Uncertainty in Artificial Intelligence*, pages 261–270. PMLR.

Xu, X. and Lyu, L. (2020). Towards building a robust and fair federated learning system. *arXiv preprint arXiv:2011.10464*.

Yin, D., Chen, Y., Kannan, R., and Bartlett, P. (2018). Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pages 5650–5659. PMLR.

Zhu, H. and Ling, Q. (2021). Broadcast: Reducing both stochastic and compression noise to robustify communication-efficient federated learning. *arXiv preprint arXiv:2104.06685*.

Zinkevich, M., Weimer, M., Li, L., and Smola, A. (2010). Parallelized stochastic gradient descent. *Advances in neural information processing systems*, 23.

Checklist

1. For all models and algorithms presented, check if you include:

(a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Contents

A EXTRA RELATED WORK

Unbiased compression. The first convergence results for the methods using unbiased compression of (stochastic) gradients are established by Alistarh et al. (2017); Wen et al. (2017); Khirirat et al. (2018). Mishchenko et al. (2019) propose an approach based on the compression of certain gradient differences, achieving convergence to arbitrary accuracy. In this method, workers compute full gradients and a constant stepsize is used. This approach is generalized and combined with variance reduction by Horváth et al. (2019b). The methods achieving current state-of-the-art theoretical complexities in the non-convex case are developed by Gorbunov et al. (2021b) and Tyurin and Richtárik (2023b). Recent advancements in the literature on methods utilising unbiased compression include adaptive compression (Faghri et al., 2020), acceleration (Li et al., 2020; Li and Richtárik, 2021; Qian et al., 2021), decentralized communication (Kovalev et al., 2021), local steps (Basu et al., 2019; Haddadpour et al., 2021; Sadiev et al., 2022), second-order methods (Islamov et al., 2021; Safaryan et al., 2021), and methods for the saddle-point problems (Beznosikov et al., 2021, 2022).

Biased compression and error feedback. Methods with biased compression and error feedback are known to perform well in practice (Seide et al., 2014; Vogels et al., 2019). In the non-convex case, which is the primary focus of our work, standard error feedback is analyzed by Karimireddy et al. (2019); Beznosikov et al. (2020); Koloskova et al. (2019); Sahu et al. (2021). However, the existing complexity bounds for standard error feedback either have an explicit dependence on the heterogeneity parameter ζ^2 or require boundedness of the gradients. These issues are resolved by Richtárik et al. (2021), who propose a novel version of error feedback, named EF21. This approach is further extended in various directions by Fatkhullin et al. (2021).

Bidirectional compression. In some applications, the communication cost of downloading information from the server is comparable to the cost of uploading it, as observed in data from Speedtest.net[4]. Consequently, numerous studies not only address uplink (workers-to-server) compression costs, but also focus on downlink (server-to-workers) compression. Philippenko and Dieuleveut (2020); Gorbunov et al. (2020) utilize unbiased compression in both uplink and downlink communication, and Philippenko and Dieuleveut (2021) improve these results by employing the DIANA mechanism (Mishchenko et al., 2019) on both worker and server sides. There also exist several works considering biased compression. Tang et al. (2019) use standard error feedback, and Fatkhullin et al. (2021) apply EF21 on both workers and server sides. The latter approach is further refined by Gruntkowska et al. (2023); Tyurin and Richtárik (2023a).

Comparison with Tao et al. (2023). Tao et al. (2023) consider a variant of Parallel SGD with coordinate-wise Truncated Mean aggregation and coordinate-wise robust estimate of the gradient on regular workers to handle the heavy-tailed noise in the stochastic gradients. Additionally, the authors apply contractive compression to messages communicated by the clients. However, their approach assumes that all clients sample from the same distribution and have a finite number of samples, while we consider a setup with bounded heterogeneity without assuming statistical similarity between local datasets on each client. Moreover, Tao et al. (2023) focus solely on strongly convex problems, deriving high-probability results. Thus, one cannot formally compare our theoretical guarantees. Importantly, Tao et al. (2023) assume that the second moments of the stochastic gradients are bounded, which implies boundedness of the gradient of the objective function. This assumption is known to be quite restrictive and, for example, does not hold for globally strongly convex functions. Finally, in the special case of homogeneous data, our analysis recovers convergence to any predefined accuracy, while theirs ensures convergence only to the (non-reducible by stepsize) neighborhood of the solution, with a radius proportional to the problem dimension, which can be huge.

[4] https://www.speedtest.net/global-index

B EXAMPLES OF ROBUST AGGREGATORS AND COMPRESSION OPERATORS

B.1 Robust Aggregators

In recent years, research efforts have given rise to various aggregation rules asserted to be Byzantine-robust under certain assumptions (Li et al., 2023). However, it turns out that these rules do not satisfy Definition 1.1, and there exist practical scenarios where methods employing them fail to converge (Karimireddy et al., 2021).

In this section, we present a few examples of such rules and consider a technique known as *bucketing*, proposed by Karimireddy et al. (2022), which robustify them.

Geometric Median (Chen et al., 2017; Pillutla et al., 2022): GM-estimator, also known as Robust Federated Averaging (RFA), is an aggregation rule based on geometric median:

$$\texttt{GM}(x_1, \ldots, x_n) \stackrel{\text{def}}{=} \operatorname*{argmin}_{x \in \mathbb{R}^d} \sum_{i=1}^{n} \|x - x_i\|.$$

An approximate solution to the above problem can be obtained through several iterations of the smoothed Weiszfeld algorithm, with $\mathcal{O}(n)$ cost per iteration (Weiszfeld, 1937; Pillutla et al., 2022).

Coordinate-wise Median (Yin et al., 2018): CM-estimator operates component-wise, assigning to each element in the output vector the median of the corresponding elements in the input vectors, i.e.,

$$[\texttt{CM}(x_1, \ldots, x_n)]_j \stackrel{\text{def}}{=} \texttt{Median}([x_1]_j, \ldots, [x_n]_j),$$

where $[x]_j$ is j-th coordinate of vector $x \in \mathbb{R}^d$. The method incurs a cost of $\mathcal{O}(n)$.

Krum (Blanchard et al., 2017b): Krum outputs a vector x_i which is closest to the mean of the input vectors when the most extreme inputs are excluded. In mathematical terms, let $S_i \subseteq \{x_1, \ldots, x_n\}$ represent the subset of $n - |\mathcal{B}| - 2$ vectors closest to x_i. Then

$$\texttt{Krum}(x_1, \ldots, x_n) \stackrel{\text{def}}{=} \operatorname*{argmin}_{x_i \in \{x_1, \ldots, x_n\}} \sum_{j \in S_i} \|x_j - x_i\|^2.$$

This estimator has a serious limitation, namely the computational cost. Since it requires calculating all pairwise distances between vectors $x_1, \ldots, x_n$, this cost is $\mathcal{O}(n^2)$.

Bucketing (Karimireddy et al., 2022): The s-bucketing trick (Algorithm 5) first divides the n inputs into $\lceil n/s \rceil$ buckets, so that each bucket contains at most s elements. The vectors in each bucket are then averaged and employed as input of some aggregator Aggr.

Algorithm 5 Robust Aggregation with Bucketing (Karimireddy et al., 2022)

Input: $\{x_1, \ldots, x_n\}$, bucket size $s \in \mathbb{N}$, aggregation rule Aggr
Sample a random permutation $\pi = (\pi(1), \ldots, \pi(n))$ of $[n]$
Set $y_i = \frac{1}{s} \sum_{k=s(i-1)+1}^{\min\{si,n\}} x_{\pi(k)}$ for $i = 1, \ldots, \lceil n/s \rceil$
Return: $\hat{x} = \texttt{Aggr}(y_1, \ldots, y_{\lceil n/s \rceil})$

Gorbunov et al. (2023) show that Krum, RFA and CM with bucketing satisfy Definition 1.1.

B.2 Compression Operators

Below we provide several examples of compression operators belonging to the classes of unbiased (Definition 1.2) and contractive (Definition 1.3) compressors. For a more thorough overview, we refer the reader to Beznosikov et al. (2020).

1. **Rand$K$ sparsification** (Stich et al., 2018): The Rand-$K \in \mathbb{U}(\frac{d}{K}-1)$ compressor retains $K \in [d]$ random values of the input vector and scales it by $\frac{d}{K}$:

$$\mathcal{C}(x) = \frac{d}{K}\sum_{i\in S} x_i e_i,$$

where S is a random subset of $[d]$ and $e_1, \ldots, e_d \in \mathbb{R}^d$ are the standard unit basis vectors.

2. **Natural compression** (Horváth et al., 2019a): The compressor is defined component-wise via $(\mathcal{C}(x))_i = \mathcal{C}(x_i)$. For $x_i \in \mathbb{R}$, we set

$$\mathcal{C}(0) \stackrel{\text{def}}{=} 0$$

and for $x_i \neq 0$

$$\mathcal{C}(x_i) \stackrel{\text{def}}{=} \begin{cases} \text{sign}(x_i)\cdot 2^{\lfloor \log_2 |x_i| \rfloor} & \text{with probability } p(x_i), \\ \text{sign}(x_i)\cdot 2^{\lceil \log_2 |x_i| \rceil} & \text{with probability } 1-p(x_i), \end{cases}$$

where

$$p(x_i) \stackrel{\text{def}}{=} \frac{2^{\lceil \log_2 |x_i| \rceil} - |x_i|}{2^{\lfloor \log_2 |x_i| \rfloor}}.$$

It can be shown that $\mathcal{C} \in \mathbb{U}(\frac{1}{8})$.

3. **Top$K$ sparsification** (Alistarh et al., 2018b): The Top-$K \in \mathbb{B}(\frac{K}{d})$ sparsifier keeps the $K \in [d]$ largest coordinates in magnitude (ordered so that $|x_{(1)}| \leq \ldots \leq |x_{(d)}|$):

$$\mathcal{C}(x) = \sum_{i=d-K+1}^{d} x_{(i)} e_{(i)}.$$

4. **Biased random sparsification** (Beznosikov et al., 2020): Let $S \subseteq [d]$ and $p_i = \mathbb{P}(i \in S) > 0$ for $i \in [d]$ and define

$$\mathcal{C}(x) = \sum_{i\in S} x_i e_i.$$

The biased random sparsification operator belongs to $\mathbb{B}(\min_i p_i)$.

5. **General biased rounding** (Beznosikov et al., 2020): Let

$$(\mathcal{C}(x))_i = \text{sign}(x_i) \arg\min_{t\in(a_k)} |t - |x_i||, \qquad i \in [d],$$

where $(a_k)_{k\in\mathbb{Z}}$ is an increasing sequence of positive numbers such that $\inf a_k = 0$ and $\sup a_k = \infty$. This operator belongs to $\mathbb{B}(\alpha)$, where $\alpha^{-1} = \sup_{k\in\mathbb{Z}} \frac{(a_k+a_{k+1})^2}{4a_k a_{k+1}}$.

C USEFUL IDENTITIES AND INEQUALITIES

For all $x, y \in \mathbb{R}^d$, $s > 0$ and $\alpha \in (0, 1]$, we have:

$$\|x + y\|^2 \leq (1 + s) \|x\|^2 + (1 + s^{-1}) \|y\|^2, \tag{8}$$

$$(1 - \alpha)\left(1 + \frac{\alpha}{2}\right) \leq 1 - \frac{\alpha}{2}, \tag{9}$$

$$(1 - \alpha)\left(1 + \frac{2}{\alpha}\right) \leq \frac{2}{\alpha}. \tag{10}$$

Jensen's inequality: If f is a convex function and X is a random variable, then

$$\mathbb{E}\left[f(X)\right] \geq f\left(\mathbb{E}\left[X\right]\right). \tag{11}$$

Variance decomposition: For any random vector $X \in \mathbb{R}^d$ and any non-random vector $c \in \mathbb{R}^d$, we have

$$\mathbb{E}\left[\|X - c\|^2\right] = \mathbb{E}\left[\|X - \mathbb{E}\left[X\right]\|^2\right] + \|\mathbb{E}\left[X\right] - c\|^2. \tag{12}$$

Tower property: For any random variables X and Y, we have

$$\mathbb{E}\left[\mathbb{E}\left[X \,|\, Y\right]\right] = \mathbb{E}\left[X\right]. \tag{13}$$

Lemma C.1 (Lemma 2 of (Li et al., 2021)). *Suppose that function f is L-smooth and let $x^{t+1} = x^t - \gamma g^t$. Then for any $g^t \in \mathbb{R}^d$ and $\gamma > 0$, 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 C.2 (Lemma 5 of (Richtárik et al., 2021)). *Let $a, b > 0$. 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 $\frac{1}{\sqrt{a}+b} \leq \min\left\{\frac{1}{\sqrt{a}}, \frac{1}{b}\right\} \leq \frac{2}{\sqrt{a}+b}$.*

D SKETCH OF THE PROOFS

The core intuition of our work centers around reducing the variance in stochastic gradients generated by reliable clients, a goal achieved through the following application of Young's inequality:

$$\begin{aligned}\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i-\nabla f(x)\right\|^2\right] &\leq 2\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i-\nabla f_i(x)\right\|^2\right]+2\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\nabla f_i(x)-\nabla f(x)\right\|^2\right]\\ &\leq 2\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i-\nabla f_i(x)\right\|^2\right]+2\zeta^2\end{aligned}$$

Effectively bounding the variance of stochastic gradients g_i requires constructing them in a way such that:

$$\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i-\nabla f_i(x)\right\|^2\right]\underset{x\to x^*}{\to}0 \quad \text{where} \quad x^*=\arg\min_{x\in\mathbb{R}^d} f(x)$$

Our algorithms are designed to ensure this convergence. In a neighborhood of x^*, we establish:

$$\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i-\nabla f(x)\right\|^2\right]\leq 2\zeta^2$$

To achieve this, we introduce an unconventional Byzantine-related descent lemma for the first time in the literature of optimization with Byzantine workers (Lemmas E.2 and F.1) And then we build on the components of the desent lemma inequality, in order to get the convergence results. Full proofs are provided in the following sections.

E MISSING PROOFS FOR Byz-VR-MARINA 2.0

For the sake of clarity, we adopt the following notation: $\overline{g}^t \overset{\text{def}}{=} \frac{1}{G}\sum_{i\in\mathcal{G}} g_i^t$, $R^t \overset{\text{def}}{=} \left\|x^{t+1}-x^t\right\|^2$, $H_i^t \overset{\text{def}}{=} \|g_i^t - \nabla f_i(x^t)\|^2$, $H^t \overset{\text{def}}{=} \frac{1}{G}\sum_{i\in\mathcal{G}} H_i^t$.

E.1 Technical Lemmas

We first prove several lemmas needed to prove the main result. The first two of them are based on the definition of a robust aggregator and are not specific to Byz-VR-MARINA 2.0 only.

Lemma E.1 (Bound on the variance)**.** *Suppose that Assumption 1.4 holds. Then*

$$\frac{1}{G(G-1)}\sum_{i,l\in\mathcal{G}} \mathbb{E}\left[\left\|g_i^t - g_l^t\right\|^2\right] \leq 8\mathbb{E}\left[H^t\right] + 8B\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + 8\zeta^2.$$

Proof. Denoting $u_i \overset{\text{def}}{=} g_i^t - \nabla f_i(x^t)$ and $v_i \overset{\text{def}}{=} \nabla f_i(x^t) - \nabla f(x^t)$, we have

$$\begin{aligned}
\frac{1}{G(G-1)}\sum_{i,l\in\mathcal{G}} \mathbb{E}\left[\|g_i^t - g_l^t\|^2\right] &= \frac{1}{G(G-1)}\sum_{\substack{i,l\in\mathcal{G}\\ i\neq l}} \mathbb{E}\left[\|(u_i+v_i)-(u_l+v_l)\|^2\right]\\
&\overset{(8)}{\leq} \frac{4}{G(G-1)}\sum_{\substack{i,l\in\mathcal{G}\\ i\neq l}} \left(\mathbb{E}\left[\|u_i\|^2\right] + \mathbb{E}\left[\|v_i\|^2\right] + \mathbb{E}\left[\|u_l\|^2\right] + \mathbb{E}\left[\|v_l\|^2\right]\right)\\
&= \frac{8}{G}\sum_{i\in\mathcal{G}} \left(\mathbb{E}\left[\|u_i\|^2\right] + \mathbb{E}\left[\|v_i\|^2\right]\right)\\
&= 8\mathbb{E}\left[H^t\right] + \frac{8}{G}\sum_{i\in\mathcal{G}} \mathbb{E}\left[\left\|\nabla f_i(x^t) - \nabla f(x^t)\right\|^2\right]\\
&\overset{(1.4)}{\leq} 8\mathbb{E}\left[H^t\right] + 8B\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + 8\zeta^2
\end{aligned}$$

as needed. □

Lemma E.2 (Descent Lemma)**.** *Suppose that Assumptions 1.1 and 1.4 hold. Then for all $s>0$ we have*

$$\begin{aligned}
\mathbb{E}\left[f(x^{t+1})\right] \leq\ & \mathbb{E}\left[f(x^t)\right] - \frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\mathbb{E}\left[R^t\right] + 4\gamma c\delta(1+s)\mathbb{E}\left[H^t\right]\\
&+\frac{\gamma}{2}(1+s^{-1})\mathbb{E}\left[\left\|\overline{g}^t - \nabla f(x^t)\right\|^2\right] + 4\gamma c\delta(1+s)\zeta^2,
\end{aligned}$$

where $\kappa = 1 - 8Bc\delta(1+s)$.

Proof. First, by Young's inequality, for any $s>0$ we have

$$\begin{aligned}
\mathbb{E}\left[\left\|g^t - \nabla f(x^t)\right\|^2\right] &\overset{(8)}{\leq} (1+s)\mathbb{E}\left[\left\|g^t - \overline{g}^t\right\|^2\right] + (1+s^{-1})\mathbb{E}\left[\left\|\overline{g}^t - \nabla f(x^t)\right\|^2\right]\\
&\overset{(2)}{\leq} (1+s)\frac{c\delta}{G(G-1)}\sum_{i,l\in\mathcal{G}} \mathbb{E}\left[\|g_i^t - g_l^t\|^2\right] + (1+s^{-1})\mathbb{E}\left[\left\|\overline{g}^t - \nabla f(x^t)\right\|^2\right]\\
&\overset{(E.1)}{\leq} 8(1+s)c\delta\mathbb{E}\left[H^t\right] + 8B(1+s)c\delta\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right]\\
&\quad +8(1+s)c\delta\zeta^2 + (1+s^{-1})\mathbb{E}\left[\left\|\overline{g}^t - \nabla f(x^t)\right\|^2\right].
\end{aligned}$$

Taking expectation in Lemma C.1 and applying the above bound, we obtain

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

$$
\begin{aligned}
\leq \quad & \mathbb{E}\left[f(x^t)\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\mathbb{E}\left[R^t\right] \\
& +4\gamma(1+s)c\delta\mathbb{E}\left[H^t\right] + 4\gamma B(1+s)c\delta\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \\
& +4\gamma(1+s)c\delta\zeta^2 + \frac{\gamma}{2}(1+s^{-1})\mathbb{E}\left[\left\|\overline{g}^t - \nabla f(x^t)\right\|^2\right] \\
= \quad & \mathbb{E}\left[f(x^t)\right] - \left(\frac{\gamma}{2} - 4\gamma B(1+s)c\delta\right)\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\mathbb{E}\left[R^t\right] \\
& +4\gamma(1+s)c\delta\mathbb{E}\left[H^t\right] + 4\gamma(1+s)c\delta\zeta^2 + \frac{\gamma}{2}(1+s^{-1})\mathbb{E}\left[\left\|\overline{g}^t - \nabla f(x^t)\right\|^2\right].
\end{aligned}
$$

Letting $\kappa = 1 - 8Bc\delta(1+s)$, we get the result. □

We next provide recursive bounds on $\mathbb{E}\left[\left\|\overline{g}^t - \nabla f(x^t)\right\|^2\right]$ and $H^t = \frac{1}{G}\sum_{i\in\mathcal{G}}\left\|g_i^t - \nabla f_i(x^t)\right\|^2$.

Lemma E.3. *Suppose that Assumptions 1.1, 1.2 and 1.3 hold. Then*

$$
\mathbb{E}\left[\left\|\overline{g}^{t+1} - \nabla f(x^{t+1})\right\|^2\right] \leq (1-p)\left(\mathbb{E}\left[\left\|\overline{g}^t - \nabla f(x^t)\right\|^2\right] + \left(\frac{\omega}{G}\left(\frac{\mathcal{L}_\pm^2}{b} + L_\pm^2 + L^2\right) + \frac{\mathcal{L}_\pm^2}{Gb}\right)\mathbb{E}\left[R^t\right]\right).
$$

Proof. First, we know that

$$
\begin{aligned}
& \mathbb{E}\left[\left\|\overline{g}^{t+1} - \nabla f(x^{t+1})\right\|^2\right] \\
& \quad = (1-p)\,\mathbb{E}\left[\left\|\overline{g}^t + \frac{1}{G}\sum_{i\in\mathcal{G}}\left(\mathcal{C}_i\left(\widehat{\Delta}_i\left(x^{t+1}, x^t\right)\right) - \nabla f_i\left(x^{t+1}\right)\right)\right\|^2\right] \\
& \quad = (1-p)\,\mathbb{E}\left[\left\|\overline{g}^t - \nabla f\left(x^t\right) + \frac{1}{G}\sum_{i\in\mathcal{G}}\left(\mathcal{C}_i\left(\widehat{\Delta}_i\left(x^{t+1}, x^t\right)\right) - \left(\nabla f_i\left(x^{t+1}\right) - \nabla f_i\left(x^t\right)\right)\right)\right\|^2\right].
\end{aligned}
$$

Denoting by $\mathbb{E}_{\mathcal{C}}\left[\cdot\right]$ the expectation taken with respect to the randomness of the compressor and using the tower property of conditional expectation, we have

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{C}_i\left(\widehat{\Delta}_i\left(x^{t+1}, x^t\right)\right)\right] &\overset{(13)}{=} \mathbb{E}\left[\mathbb{E}_{\mathcal{C}}\left[\mathcal{C}_i\left(\widehat{\Delta}_i\left(x^{t+1}, x^t\right)\right)\right]\right] \\
&= \mathbb{E}\left[\widehat{\Delta}_i\left(x^{t+1}, x^t\right)\right] \\
&= \mathbb{E}\left[\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right].
\end{aligned}
$$

Hence, using the independence of compressors $\{\mathcal{C}_i\}_{i\in\mathcal{G}}$ and estimators $\{\widehat{\Delta}_i\left(x^{t+1}, x^t\right)\}_{i\in\mathcal{G}}$ gives

$$
\begin{aligned}
& \mathbb{E}\left[\left\|\overline{g}^{t+1} - \nabla f(x^{t+1})\right\|^2\right] \\
& \overset{(12)}{=} (1-p)\left(\mathbb{E}\left[\left\|\overline{g}^t - \nabla f(x^t)\right\|^2\right] + \frac{1}{G^2}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\mathcal{C}_i\left(\widehat{\Delta}_i\left(x^{t+1}, x^t\right)\right) - \left(\nabla f_i\left(x^{t+1}\right) - \nabla f_i\left(x^t\right)\right)\right\|^2\right]\right) \\
& \overset{(12)}{=} (1-p)\left(\mathbb{E}\left[\left\|\overline{g}^t - \nabla f(x^t)\right\|^2\right]\right. \\
& \quad \left. + \frac{1}{G^2}\sum_{i\in\mathcal{G}}\left(\mathbb{E}\left[\left\|\mathcal{C}_i\left(\widehat{\Delta}_i\left(x^{t+1}, x^t\right)\right) - \widehat{\Delta}_i\left(x^{t+1}, x^t\right)\right\|^2\right] + \mathbb{E}\left[\left\|\widehat{\Delta}_i\left(x^{t+1}, x^t\right) - \Delta_i\left(x^{t+1}, x^t\right)\right\|^2\right]\right)\right) \\
& \overset{(1.3)}{\leq} (1-p)\left(\mathbb{E}\left[\left\|\overline{g}^t - \nabla f(x^t)\right\|^2\right] + \frac{\omega}{G^2}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\widehat{\Delta}_i\left(x^{t+1}, x^t\right)\right\|^2\right] + \frac{\mathcal{L}_\pm^2}{Gb}\mathbb{E}\left[\left\|x^{t+1} - x^t\right\|^2\right]\right) \\
& \overset{(12)}{=} (1-p)\,\mathbb{E}\left[\left\|\overline{g}^t - \nabla f(x^t)\right\|^2\right] \\
& \quad + (1-p)\,\frac{\omega}{G^2}\sum_{i\in\mathcal{G}}\left(\mathbb{E}\left[\left\|\widehat{\Delta}_i\left(x^{t+1}, x^t\right) - \Delta_i\left(x^{t+1}, x^t\right)\right\|^2\right] + \mathbb{E}\left[\left\|\Delta_i\left(x^{t+1}, x^t\right)\right\|^2\right]\right)
\end{aligned}
$$

$$
\begin{aligned}
&+ (1-p)\frac{\mathcal{L}_{\pm}^2}{Gb}\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right] \\
&\overset{(1.2),(1.3)}{\leq} (1-p)\,\mathbb{E}\left[\left\|\overline{g}^t-\nabla f(x^t)\right\|^2\right] + (1-p)\frac{\omega}{G}\left(\frac{\mathcal{L}_{\pm}^2}{b}+L_{\pm}^2+L^2\right)\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right] \\
&+ (1-p)\frac{\mathcal{L}_{\pm}^2}{Gb}\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right] \\
&= (1-p)\left(\mathbb{E}\left[\left\|\overline{g}^t-\nabla f(x^t)\right\|^2\right] + \left(\frac{\omega}{G}\left(\frac{\mathcal{L}_{\pm}^2}{b}+L_{\pm}^2+L^2\right)+\frac{\mathcal{L}_{\pm}^2}{Gb}\right)\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right]\right).
\end{aligned}
$$

□

Lemma E.4. *Suppose that Assumptions 1.1, 1.2 and 1.3 hold. Then*

$$
\mathbb{E}\left[H^{t+1}\right] \qu\leq \quad (1-p)\left(\mathbb{E}\left[H^t\right] + \left(\omega\left(\frac{\mathcal{L}_{\pm}^2}{b}+L_{\pm}^2+L^2\right)+\frac{\mathcal{L}_{\pm}^2}{b}\right)\mathbb{E}\left[R^t\right]\right).
$$

Proof. Following the same reasoning as in the proof of the previous lemma gives

$$
\begin{aligned}
\mathbb{E}\left[H_i^{t+1}\right] &= \mathbb{E}\left[\left\|g_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right] \\
&= (1-p)\,\mathbb{E}\left[\left\|g_i^t+\mathcal{C}_i\left(\widehat{\Delta}_i\left(x^{t+1},x^t\right)\right)-\nabla f_i\left(x^{t+1}\right)\right\|^2\right] \\
&= (1-p)\,\mathbb{E}\left[\left\|g_i^t-\nabla f_i(x^t)+\mathcal{C}_i\left(\widehat{\Delta}_i\left(x^{t+1},x^t\right)\right)-\left(\nabla f_i\left(x^{t+1}\right)-\nabla f_i\left(x^t\right)\right)\right\|^2\right] \\
&\overset{(12)}{=} (1-p)\left(\mathbb{E}\left[\left\|g_i^t-\nabla f_i(x^t)\right\|^2\right]+\mathbb{E}\left[\left\|\mathcal{C}_i\left(\widehat{\Delta}_i\left(x^{t+1},x^t\right)\right)-\left(\nabla f_i\left(x^{t+1}\right)-\nabla f_i\left(x^t\right)\right)\right\|^2\right]\right).
\end{aligned}
$$

Averaging over all the good workers $i\in\mathcal{G}$ and using Assumptions 1.2 and 1.3 gives the result. □

E.2 Proof of Theorem 2.1

For the readers' convenience, we repeat the statement of the theorem.

Theorem 2.1. *Let Assumptions 1.1, 1.2, 1.3 and 1.4 hold. Assume that* $0<\gamma\leq(L+\sqrt{\eta})^{-1}$, $\delta<((8c+4\sqrt{c})B)^{-1}$ *and initialize* $g_i^0=\nabla f_i(x^0)$ *for all* $i\in\mathcal{G}$, *where* $\eta=\frac{1-p}{p}\left(\omega\left(\frac{\mathcal{L}_{\pm}^2}{b}+L_{\pm}^2+L^2\right)+\frac{\mathcal{L}_{\pm}^2}{b}\right)\left(\sqrt{\frac{1}{G}}+\sqrt{8c\delta}\right)^2$. *Then for all* $T\geq 0$ *the iterates produced by* Byz-VR-MARINA 2.0 *satisfy*

$$
\mathbb{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right]\leq\frac{1}{A}\left(\frac{2\delta^0}{\gamma T}+\left(8c\delta+\sqrt{\frac{8c\delta}{G}}\right)\zeta^2\right),
$$

where $\delta^0=f(x^0)-f^*$, $A=1-\left(8c\delta+\sqrt{8c\delta/G}\right)B$ *and* $\widehat{x}^T$ *is chosen uniformly at random from* $x^0,x^1,\ldots,x^{T-1}$.

Proof. Let $C=\frac{1+s^{-1}}{p}$, $D=\frac{8c\delta(1+s)}{p}$ for some $s>0$ and define

$$
\begin{aligned}
\psi_t &= C\left\|\overline{g}^t-\nabla f(x^t)\right\|^2+DH^t, \\
\Phi_t &= f(x^t)-f^*+\frac{\gamma}{2}\psi_t.
\end{aligned}
$$

Using lemmas E.2, E.3 and E.4, we have

$$
\begin{aligned}
\mathbb{E}\left[\Phi_{t+1}\right] &= \mathbb{E}\left[f(x^{t+1})-f^*+\frac{\gamma}{2}\psi_{t+1}\right] \\
&= \underbrace{\mathbb{E}\left[f(x^{t+1})-f^*\right]}_{\text{use E.2}}+\frac{\gamma}{2}\left(C\underbrace{\mathbb{E}\left[\left\|\overline{g}^{t+1}-\nabla f(x^{t+1})\right\|^2\right]}_{\text{use E.3}}+D\underbrace{\mathbb{E}\left[H^{t+1}\right]}_{\text{use E.4}}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[f(x^t)-f^*\right]+\frac{\gamma}{2}\left(\underbrace{C\mathbb{E}\left[\left\|\bar{g}^t-\nabla f\left(x^t\right)\right\|^2\right]+D\mathbb{E}\left[H^t\right]}_{\mathbb{E}\left[\psi_t\right]}\right)-\frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right]\\
&\quad-\left(\frac{1}{2\gamma}-\frac{L}{2}-\frac{\gamma\eta}{2}\right)\mathbb{E}\left[R^t\right]+4\gamma c\delta(1+s)\zeta^2\\
&= \mathbb{E}\left[\Phi_t\right]-\frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right]-\left(\frac{1}{2\gamma}-\frac{L}{2}-\frac{\gamma\eta}{2}\right)\mathbb{E}\left[R^t\right]+4\gamma c\delta(1+s)\zeta^2,
\end{aligned}
$$

where

$$
\begin{aligned}
\eta \overset{\text{def}}{=} & C\left(1-p\right)\left(\frac{\omega}{G}\left(\frac{\mathcal{L}_{\pm}^2}{b}+L_{\pm}^2+L^2\right)+\frac{\mathcal{L}_{\pm}^2}{Gb}\right)+D\left(1-p\right)\left(\omega\left(\frac{\mathcal{L}_{\pm}^2}{b}+L_{\pm}^2+L^2\right)+\frac{\mathcal{L}_{\pm}^2}{b}\right)\\
=&\left(1-p\right)\left(\omega\left(\frac{\mathcal{L}_{\pm}^2}{b}+L_{\pm}^2+L^2\right)+\frac{\mathcal{L}_{\pm}^2}{b}\right)\left(\frac{C}{G}+D\right)\\
=&\frac{1-p}{p}\left(\omega\left(\frac{\mathcal{L}_{\pm}^2}{b}+L_{\pm}^2+L^2\right)+\frac{\mathcal{L}_{\pm}^2}{b}\right)\left(\frac{1+s^{-1}}{G}+8c\delta\left(1+s\right)\right).
\end{aligned}
$$

Taking $0<\gamma\leq\frac{1}{L+\sqrt{\eta}}$ and using Lemma C.2 gives

$$
\frac{1}{2\gamma}-\frac{L}{2}-\frac{\gamma\eta}{2}\geq 0
$$

and hence

$$
\mathbb{E}\left[\Phi_{t+1}\right]\leq\mathbb{E}\left[\Phi_t\right]-\frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right]+4\gamma c\delta(1+s)\zeta^2.
$$

Now, let us take $s=\frac{1}{\sqrt{8c\delta G}}$. Recalling that $\kappa=1-8Bc\delta(1+s)$, our assumption on δ implies that $\kappa>0$ and hence, summing up the above inequality for $t=0,1,\ldots,T-1$ and rearranging the terms, we get

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right]&\leq\frac{2}{\gamma\kappa T}\left(\mathbb{E}\left[\Phi_0\right]-\mathbb{E}\left[\Phi_T\right]\right)+\frac{8c\delta(1+s)\zeta^2}{\kappa}\\
&\leq\frac{1}{\kappa}\left(\frac{2}{\gamma T}\mathbb{E}\left[\Phi_0\right]+8c\delta(1+s)\zeta^2\right)\\
&=\frac{1}{1-\left(8c\delta+\sqrt{\frac{8c\delta}{G}}\right)B}\left(\frac{2\left(f(x^0)-f^*\right)}{\gamma T}+\left(8c\delta+\sqrt{\frac{8c\delta}{G}}\right)\zeta^2\right)
\end{aligned}
$$

as needed. □

Corollary E.1. *Let the assumptions of Theorem 2.1 hold, $p=\min\{1/\omega,b/m\}$ and $\gamma=\left(L+\sqrt{\eta}\right)^{-1}$, where $\eta=\max\{\omega,m/b\}\left(\omega\left(\frac{\mathcal{L}_{\pm}^2}{b}+L_{\pm}^2+L^2\right)+\frac{\mathcal{L}_{\pm}^2}{b}\right)\left(\sqrt{\frac{1}{G}}+\sqrt{8c\delta}\right)^2$. Then for all $T\geq 0$, $\mathbb{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right]$ is of the order*

$$
\mathcal{O}\left(\left(\frac{\left(L+\sqrt{\max\{\omega^2,m\omega/b\}\left(\frac{\mathcal{L}_{\pm}^2}{b}+L_{\pm}^2+L^2\right)}\left(\sqrt{\frac{1}{G}}+\sqrt{c\delta}\right)\right)\delta^0}{T\left(1-\left(c\delta+\sqrt{\frac{c\delta}{G}}\right)B\right)}+\frac{\left(c\delta+\sqrt{\frac{c\delta}{G}}\right)\zeta^2}{1-\left(c\delta+\sqrt{\frac{c\delta}{G}}\right)B}\right)\right),
$$

where $\delta^0=f(x^0)-f^*$ and $\widehat{x}^T$ is chosen uniformly at random from $x^0,x^1,\ldots,x^{T-1}$. Hence, to guarantee $\mathbb{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right]\leq\varepsilon^2$ for $\varepsilon^2>\left(1-\left(8c\delta+\sqrt{8c\delta/G}\right)B\right)^{-1}\left(8c\delta+\sqrt{8c\delta/G}\right)\zeta^2$, Byz-VR-MARINA 2.0 *requires*

$$
\mathcal{O}\left(\frac{1}{\varepsilon^2}\left(1+\sqrt{\max\left\{\omega^2,\frac{m\omega}{b}\right\}}\left(\sqrt{\frac{1}{G}}+\sqrt{c\delta}\right)\right)\right)
$$

communication rounds.

F MISSING PROOFS FOR Byz-DASHA-PAGE

In this section, we again use the notation $\overline{g}^t \overset{\text{def}}{=} \frac{1}{G}\sum_{i\in\mathcal{G}} g_i^t$, $R^t \overset{\text{def}}{=} \left\|x^{t+1}-x^t\right\|^2$, $H_i^t \overset{\text{def}}{=} \left\|g_i^t - \nabla f_i(x^t)\right\|^2$, $H^t \overset{\text{def}}{=} \frac{1}{G}\sum_{i\in\mathcal{G}} H_i^t$.

F.1 Technical Lemmas

Lemma F.1 (Descent Lemma)**.** *Suppose that Assumptions 1.1 and 1.4 hold. Then for all $s>0$ we have*

$$\begin{aligned}\mathbb{E}\left[f(x^{t+1})\right] \;\leq\; & \mathbb{E}\left[f(x^t)\right] - \frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] - \left(\frac{1}{2\gamma}-\frac{L}{2}\right)\mathbb{E}\left[R^t\right] \\ & +8\gamma c\delta(1+s)\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^t-h_i^t\right\|^2\right]\right) + 8\gamma c\delta(1+s)\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]\right) \\ & +\gamma(1+s^{-1})\mathbb{E}\left[\left\|\overline{g}^t-h^t\right\|^2\right] + \gamma(1+s^{-1})\mathbb{E}\left[\left\|h^t-\nabla f(x^t)\right\|^2\right] + 4\gamma c\delta(1+s)\zeta^2,\end{aligned}$$

where $\kappa = 1-8Bc\delta(1+s)$.

Proof. Using Lemma E.2, we get

$$\begin{aligned}\mathbb{E}\left[f(x^{t+1})\right] \;\leq\; & \mathbb{E}\left[f(x^t)\right] - \frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] - \left(\frac{1}{2\gamma}-\frac{L}{2}\right)\mathbb{E}\left[R^t\right] \\ & +4\gamma c\delta(1+s)\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^t-\nabla f_i(x^t)\right\|^2\right]\right) \\ & +\frac{\gamma}{2}(1+s^{-1})\mathbb{E}\left[\left\|\overline{g}^t-\nabla f(x^t)\right\|^2\right] + 4\gamma c\delta(1+s)\zeta^2 \\ \overset{(8)}{\leq}\; & \mathbb{E}\left[f(x^t)\right] - \frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] - \left(\frac{1}{2\gamma}-\frac{L}{2}\right)\mathbb{E}\left[R^t\right] \\ & +8\gamma c\delta(1+s)\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^t-h_i^t\right\|^2\right]\right) + 8\gamma c\delta(1+s)\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]\right) \\ & +\gamma(1+s^{-1})\mathbb{E}\left[\left\|\overline{g}^t-h^t\right\|^2\right] + \gamma(1+s^{-1})\mathbb{E}\left[\left\|h^t-\nabla f(x^t)\right\|^2\right] + 4\gamma c\delta(1+s)\zeta^2.\end{aligned}$$

□

In what follows, we denote by $\mathbb{E}_{\mathcal{C}}[\cdot]$ the expectation taken with respect to the randomness of the compressor, and by $\mathbb{E}_h[\cdot]$ the expectation taken with respect to the choice of $\{h_i\}_{i\in\mathcal{G}}$.

Lemma F.2. *Suppose that Assumptions 1.1, 1.2 and 1.3 hold. Then*

$$\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^{t+1}-h_i^t\right\|^2\right] \leq \left(\frac{(1-p)\mathcal{L}_{\pm}^2}{b} + 2\left(L_{\pm}^2+L^2\right)\right)\mathbb{E}\left[R^t\right] + 2p\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]\right).$$

Proof. First, we have

$$\begin{aligned}\mathbb{E}_h\left[\left\|h_i^{t+1}-h_i^t\right\|^2\right] \;=\; & p\left\|\nabla f_i(x^{t+1})-h_i^t\right\|^2 + (1-p)\mathbb{E}_h\left[\left\|\widehat{\Delta}_i(x^{t+1},x^t)\right\|^2\right] \\ \overset{(12)}{=}\; & p\left\|\nabla f_i(x^{t+1})-h_i^t\right\|^2 + (1-p)\mathbb{E}_h\left[\left\|\widehat{\Delta}_i(x^{t+1},x^t)-\Delta_i(x^{t+1},x^t)\right\|^2\right] \\ & +(1-p)\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2 \\ \overset{(8)}{\leq}\; & 2p\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2 + 2p\left\|h_i^t-\nabla f_i(x^t)\right\|^2 \\ & +(1-p)\mathbb{E}_h\left[\left\|\widehat{\Delta}_i(x^{t+1},x^t)-\Delta_i(x^{t+1},x^t)\right\|^2\right]\end{aligned}$$

$$
\begin{aligned}
&+(1-p)\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2\\
\leq\ & 2p\left\|h_i^t-\nabla f_i(x^t)\right\|^2+(1-p)\mathbb{E}_h\left[\left\|\widehat{\Delta}_i(x^{t+1},x^t)-\Delta_i(x^{t+1},x^t)\right\|^2\right]\\
&+2\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2.
\end{aligned}
$$

Taking full expectation and averaging over all the good workers, we get

$$
\begin{aligned}
\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^{t+1}-h_i^t\right\|^2\right] \leq\ & (1-p)\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\widehat{\Delta}_i(x^{t+1},x^t)-\Delta_i(x^{t+1},x^t)\right\|^2\right]\\
&+\frac{2}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|\right]^2+\frac{2p}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^t-\nabla f_i(x^t)\right\|\right]^2.
\end{aligned}
$$

Using Assumptions 1.1, 1.2 and 1.3, we can conclude that

$$
\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^{t+1}-h_i^t\right\|^2\right]\leq\left(\frac{(1-p)\mathcal{L}_{\pm}^2}{b}+2\left(L_{\pm}^2+L^2\right)\right)\mathbb{E}\left[R^t\right]+2p\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]\right).
$$

□

Lemma F.3. *Suppose that Assumptions 1.1, 1.2 and 1.3 hold. Then*

$$
\begin{aligned}
\mathbb{E}\left[\left\|\overline{g}^{t+1}-h^{t+1}\right\|^2\right]\leq\ & (1-a)^2\,\mathbb{E}\left[\left\|\overline{g}^t-h^t\right\|^2\right]+\frac{2\omega}{G}\left(\frac{(1-p)\mathcal{L}_{\pm}^2}{b}+2\left(L_{\pm}^2+L^2\right)\right)\mathbb{E}\left[R^t\right]\\
&+\frac{4p\omega}{G}\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]\right)+\frac{2a^2\omega}{G}\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^t-h_i^t\right\|^2\right]\right).
\end{aligned}
$$

Proof. First, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{C}}\left[\left\|\overline{g}^{t+1}-h^{t+1}\right\|^2\right]\\
&=\mathbb{E}_{\mathcal{C}}\left[\left\|\overline{g}^t+\frac{1}{G}\sum_{i\in\mathcal{G}}\mathcal{C}_i\left(h_i^{t+1}-h_i^t-a\left(g_i^t-h_i^t\right)\right)-h^{t+1}\right\|^2\right]\\
&=\mathbb{E}_{\mathcal{C}}\left[\left\|\frac{1}{G}\sum_{i\in\mathcal{G}}\mathcal{C}_i\left(h_i^{t+1}-h_i^t-a\left(g_i^t-h_i^t\right)\right)-\frac{1}{G}\sum_{i\in\mathcal{G}}\left(h_i^{t+1}-h_i^t-a\left(g_i^t-h_i^t\right)\right)\right\|^2\right]\\
&\quad+(1-a)^2\left\|\overline{g}^t-h^t\right\|^2.
\end{aligned}
$$

Using the independence of compressors, we get

$$
\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[\left\|\overline{g}^{t+1}-h^{t+1}\right\|^2\right] &= \frac{1}{G^2}\sum_{i\in\mathcal{G}}\mathbb{E}_{\mathcal{C}}\left[\left\|\mathcal{C}_i\left(h_i^{t+1}-h_i^t-a\left(g_i^t-h_i^t\right)\right)-\left(h_i^{t+1}-h_i^t-a\left(g_i^t-h_i^t\right)\right)\right\|^2\right]\\
&\quad+(1-a)^2\left\|\overline{g}^t-h^t\right\|^2\\
&\leq \frac{\omega}{G^2}\sum_{i\in\mathcal{G}}\left\|h_i^{t+1}-h_i^t-a\left(g_i^t-h_i^t\right)\right\|^2+(1-a)^2\left\|\overline{g}^t-h^t\right\|^2\\
&\overset{(8)}{\leq} \frac{2\omega}{G^2}\sum_{i\in\mathcal{G}}\left\|h_i^{t+1}-h_i^t\right\|^2+\frac{2a^2\omega}{G^2}\sum_{i\in\mathcal{G}}\left\|g_i^t-h_i^t\right\|^2+(1-a)^2\left\|\overline{g}^t-h^t\right\|^2.
\end{aligned}
$$

Taking full expectation and using Lemma F.2 gives

$$
\mathbb{E}\left[\left\|\overline{g}^{t+1}-h^{t+1}\right\|^2\right] \leq \frac{2\omega}{G}\left(\frac{(1-p)\mathcal{L}_{\pm}^2}{b}+2\left(L_{\pm}^2+L^2\right)\right)\mathbb{E}\left[R^t\right]
$$

$$+\frac{4p\omega}{G}\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]\right)$$
$$+\frac{2a^2\omega}{G}\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^t-h_i^t\right\|^2\right]\right)+(1-a)^2\,\mathbb{E}\left[\left\|\overline{g}^t-h^t\right\|^2\right]$$

as needed. □

Lemma F.4. *Suppose that Assumptions 1.1, 1.2 and 1.3 hold. Then*

$$\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^{t+1}-h_i^{t+1}\right\|^2\right]\le\left(2a^2\omega+(1-a)^2\right)\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^t-h_i^t\right\|^2\right]+4p\omega\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]$$
$$+2\omega\left(\frac{(1-p)\mathcal{L}_{\pm}^2}{b}+2\left(L_{\pm}^2+L^2\right)\right)\mathbb{E}\left[R^t\right].$$

Proof. First, we use the definition of compression operators to bound $\mathbb{E}_{\mathcal{C}}\left[\left\|g_i^{t+1}-h_i^{t+1}\right\|^2\right]$:

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[\left\|g_i^{t+1}-h_i^{t+1}\right\|^2\right] &= \mathbb{E}_{\mathcal{C}}\left[\left\|g_i^t+\mathcal{C}_i\left(h_i^{t+1}-h_i^t-a\left(g_i^t-h_i^t\right)\right)-h_i^{t+1}\right\|^2\right]\\
&\overset{(12)}{\le} \mathbb{E}_{\mathcal{C}}\left[\left\|\mathcal{C}_i\left(h_i^{t+1}-h_i^t-a\left(g_i^t-h_i^t\right)\right)-\left(h_i^{t+1}-h_i^t-a\left(g_i^t-h_i^t\right)\right)\right\|^2\right]\\
&\quad+(1-a)^2\left\|g_i^t-h_i^t\right\|^2\\
&\le \omega\left\|h_i^{t+1}-h_i^t-a\left(g_i^t-h_i^t\right)\right\|^2+(1-a)^2\left\|g_i^t-h_i^t\right\|^2\\
&\overset{(8)}{\le} 2\omega\left\|h_i^{t+1}-h_i^t\right\|^2+2a^2\omega\left\|g_i^t-h_i^t\right\|^2+(1-a)^2\left\|g_i^t-h_i^t\right\|^2\\
&= 2\omega\left\|h_i^{t+1}-h_i^t\right\|^2+\left(2a^2\omega+(1-a)^2\right)\left\|g_i^t-h_i^t\right\|^2.
\end{aligned}$$

Taking full expectation and averaging over all $i\in\mathcal{G}$, we get

$$\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^{t+1}-h_i^{t+1}\right\|^2\right]\le 2\omega\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^{t+1}-h_i^t\right\|^2\right]+\left(2a^2\omega+(1-a)^2\right)\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^t-h_i^t\right\|^2\right].$$

Using Lemma F.2, we can conclude that

$$\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^{t+1}-h_i^{t+1}\right\|^2\right]\le 2\omega\left(\frac{(1-p)\mathcal{L}_{\pm}^2}{b}+2\left(L_{\pm}^2+L^2\right)\right)\mathbb{E}\left[R^t\right]+4p\omega\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]$$
$$+\left(2a^2\omega+(1-a)^2\right)\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^t-h_i^t\right\|^2\right].$$

□

Lemma F.5. *Suppose that Assumption 1.3 holds. Then*

$$\mathbb{E}\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^2\right]\le(1-p)\,\mathbb{E}\left[\left\|h^t-\nabla f(x^t)\right\|^2\right]+\frac{(1-p)\,\mathcal{L}_{\pm}^2}{Gb}\mathbb{E}\left[R^t\right].$$

Proof. Using the definition of h^{t+1}, we obtain

$$\begin{aligned}
&\mathbb{E}_h\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^2\right]\\
&=(1-p)\,\mathbb{E}_h\left[\left\|h^t+\frac{1}{G}\sum_{i\in\mathcal{G}}\widehat{\Delta}_i(x^{t+1},x^t)-\nabla f(x^{t+1})\right\|^2\right]\\
&=(1-p)\,\mathbb{E}_h\left[\left\|\left(h^t-\nabla f(x^t)\right)+\frac{1}{G}\sum_{i\in\mathcal{G}}\left(\widehat{\Delta}_i(x^{t+1},x^t)-\Delta_i(x^{t+1},x^t)\right)\right\|^2\right]
\end{aligned}$$

$$\overset{(12)}{=} (1-p)\,\mathbb{E}_h\left[\left\|\frac{1}{G}\sum_{i\in\mathcal{G}}\left(\widehat{\Delta}_i(x^{t+1},x^t)-\Delta_i(x^{t+1},x^t)\right)\right\|^2\right]+(1-p)\left\|h^t-\nabla f(x^t)\right\|^2.$$

From the unbiasedness and independence of mini-batch estimators, we get

$$\begin{aligned}
&\mathbb{E}_h\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^2\right]\\
&\quad\le\frac{(1-p)}{G^2}\sum_{i\in\mathcal{G}}\mathbb{E}_h\left[\left\|\widehat{\Delta}_i(x^{t+1},x^t)-\Delta_i(x^{t+1},x^t)\right\|^2\right]+(1-p)\left\|h^t-\nabla f(x^t)\right\|^2\\
&\quad\overset{(1.3)}{\le}\frac{(1-p)\,\mathcal{L}_{\pm}^2}{Gb}\left\|x^{t+1}-x^t\right\|^2+(1-p)\left\|h^t-\nabla f(x^t)\right\|^2.
\end{aligned}$$

Taking full expectation gives the result. □

Lemma F.6. *Suppose that Assumption 1.3 holds. Then*

$$\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\le(1-p)\,\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]+\frac{(1-p)\,\mathcal{L}_{\pm}^2}{b}\mathbb{E}\left[R^t\right].$$

Proof. Using the same reasoning as in the proof of the previous lemma, we have

$$\begin{aligned}
&\mathbb{E}_h\left[\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\\
&\quad=(1-p)\,\mathbb{E}_h\left[\left\|h_i^t+\widehat{\Delta}_i(x^{t+1},x^t)-\nabla f_i(x^{t+1})\right\|^2\right]\\
&\quad=(1-p)\,\mathbb{E}_h\left[\left\|h_i^t-\nabla f_i(x^t)+\widehat{\Delta}_i(x^{t+1},x^t)-\Delta_i(x^{t+1},x^t)\right\|^2\right]\\
&\quad\overset{(12)}{=}(1-p)\,\mathbb{E}_h\left[\left\|\widehat{\Delta}_i(x^{t+1},x^t)-\Delta_i(x^{t+1},x^t)\right\|^2\right]+(1-p)\left\|h_i^t-\nabla f_i(x^t)\right\|^2.
\end{aligned}$$

Taking full expectation and calculating the average over all $i\in\mathcal{G}$, we get

$$\begin{aligned}
&\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\\
&\quad=(1-p)\,\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\widehat{\Delta}_i(x^{t+1},x^t)-\Delta_i(x^{t+1},x^t)\right\|^2\right]+(1-p)\,\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]\\
&\quad\overset{(1.3)}{\le}\frac{(1-p)\,\mathcal{L}_{\pm}^2}{b}\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right]+(1-p)\,\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right].
\end{aligned}$$

□

The proof of the main result for Byz-DASHA-PAGE will require us to solve the following system of equations:

Lemma F.7 (Calculations)**.** *Assuming that $0<a<\frac{2}{2\omega+1}$ and $C,D,E,F\ge 0$, the solution of the system of equations*

$$\begin{cases}
C=(1-a)^2C+2\left(1+s^{-1}\right)\\
D=\left(2a^2\omega+(1-a)^2\right)D+\frac{2a^2\omega}{G}C+16c\delta\,(1+s)\\
E=(1-p)\,E+2\left(1+s^{-1}\right)\\
F=(1-p)\,F+4p\omega\left(\frac{C}{G}+D\right)+16c\delta\,(1+s)
\end{cases}$$

is

$$\begin{cases}
C=2\frac{1+s^{-1}}{1-(1-a)^2}\\
D=\frac{1}{1-2\omega a^2-(1-a)^2}\left(16\,(1+s)\,c\delta+\frac{2\omega a^2}{G}C\right)\\
E=2\frac{1+s^{-1}}{p}\\
F=4\omega\left(\frac{C}{G}+D\right)+\frac{16c\delta}{p}\,(1+s).
\end{cases}\tag{14}$$

Moreover,

$$\frac{C}{G}+D=\frac{2}{1-2\omega a^2-(1-a)^2}\left(\frac{1+s^{-1}}{G}+(1+s)\,8c\delta\right)$$
$$\frac{E}{G}+F=\left(\frac{8\omega}{1-2\omega a^2-(1-a)^2}+\frac{2}{p}\right)\left(\frac{1+s^{-1}}{G}+(1+s)\,8c\delta\right).$$

Proof. The fact that that this choice of C, D, E, F satisfies the equations can easily be verified by direct substitution of the values in (14). □

F.2 Proof of Theorem 2.2

For the readers' convenience, we repeat the statement of the theorem.

Theorem 2.2. *Let Assumptions 1.1, 1.2, 1.3 and 1.4 hold. Assume that* $0<\gamma\leq(L+\sqrt{\eta})^{-1}$, $\delta<((8c+4\sqrt{c})B)^{-1}$ *and initialize* $g_i^0=\nabla f_i(x^0)$ *for all* $i\in\mathcal{G}$, *where* $\eta=\left(8\omega(2\omega+1)\left(L_\pm^2+L^2\right)+\frac{1-p}{b}\left(12\omega\left(2\omega+1\right)+\frac{2}{p}\right)\mathcal{L}_\pm^2\right)\times\left(\sqrt{\frac{1}{G}}+\sqrt{8c\delta}\right)^2$. *Then for all* $T\geq 0$ *the iterates produced by* Byz-DASHA-PAGE *satisfy*

$$\mathbb{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right]\leq\frac{1}{A}\left(\frac{2\delta^0}{\gamma T}+\left(8c\delta+\sqrt{\frac{8c\delta}{G}}\right)\zeta^2\right),$$

where $\delta^0=f(x^0)-f^*$, $A=1-\left(8c\delta+\sqrt{8c\delta/G}\right)B$ *and* $\widehat{x}^T$ *is chosen uniformly at random from* $x^0,x^1,\ldots,x^{T-1}$.

Proof. Let

$$C=\frac{2(1+s^{-1})}{1-(1-a)^2}$$
$$D=\frac{1}{1-2\omega a^2-(1-a)^2}\left(16\left(1+s\right)c\delta+\frac{2\omega a^2}{G}C\right)$$
$$E=\frac{2}{p}(1+s^{-1})$$
$$F=4\omega\left(\frac{C}{G}+D\right)+\frac{16c\delta}{p}\left(1+s\right)$$

and

$$\psi_t=C\left\|\overline{g}^t-h^t\right\|^2+D\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\left\|g_i^t-h_i^t\right\|^2\right)+E\left\|h^t-\nabla f(x^t)\right\|^2+F\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right),$$
$$\Phi_t=f(x^t)-f^*+\frac{\gamma}{2}\psi_t.$$

Using Lemmas F.1, F.3, F.4, F.5 and F.6, we have

$$\begin{aligned}
&\mathbb{E}\left[\Phi_{t+1}\right]\\
&=\mathbb{E}\left[f(x^{t+1})-f^*+\frac{\gamma}{2}\psi_{t+1}\right]\\
&=\underbrace{\mathbb{E}\left[f(x^{t+1})-f^*\right]}_{\text{use F.1}}+\frac{\gamma}{2}C\underbrace{\mathbb{E}\left[\left\|\overline{g}^{t+1}-h^{t+1}\right\|^2\right]}_{\text{use F.3}}+\frac{\gamma}{2}D\underbrace{\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^{t+1}-h_i^{t+1}\right\|^2\right]\right)}_{\text{use F.4}}\\
&+\frac{\gamma}{2}E\underbrace{\mathbb{E}\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^2\right]}_{\text{use F.5}}+\frac{\gamma}{2}F\underbrace{\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\right)}_{\text{use F.6}}
\end{aligned}$$

$$\leq \mathbb{E}\left[f(x^t) - f^*\right] - \frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] - \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{\gamma\eta}{2}\right)\mathbb{E}\left[R^t\right] + 4\gamma c\delta(1+s)\zeta^2$$

$$+ \frac{\gamma}{2}\Bigg(\underbrace{C\mathbb{E}\left[\left\|\overline{g}_t - h^t\right\|^2\right] + D\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^t - h_i^t\right\|^2\right]\right) + E\mathbb{E}\left[\left\|h^t - \nabla f(x^t)\right\|^2\right] + F\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^t - \nabla f_i(x^t)\right\|^2\right]\right)}_{\mathbb{E}[\psi_t]}\Bigg)$$

$$= \mathbb{E}\left[f(x^t) - f^*\right] + \frac{\gamma}{2}\mathbb{E}\left[\psi_t\right] - \frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] - \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{\gamma\eta}{2}\right)\mathbb{E}\left[R^t\right] + 4\gamma c\delta(1+s)\zeta^2,$$

where the last inequality follows from Lemma F.7 and

$$\eta \overset{\text{def}}{=} 2\omega\left(\frac{1-p}{b}\mathcal{L}_\pm^2 + 2\left(L_\pm^2 + L^2\right)\right)\left(\frac{C}{G} + D\right) + \frac{1-p}{b}\mathcal{L}_\pm^2\left(\frac{E}{G} + F\right)$$
$$= \left(\frac{8\omega\left(L_\pm^2 + L^2\right)}{1 - 2\omega a^2 - (1-a)^2} + \frac{1-p}{b}\mathcal{L}_\pm^2\left(\frac{12\omega}{1 - 2\omega a^2 - (1-a)^2} + \frac{2}{p}\right)\right)\left(\frac{1+s^{-1}}{G} + (1+s)\,8c\delta\right).$$

Taking $0 < \gamma \leq \frac{1}{L+\sqrt{\eta}}$ and using Lemma C.2 gives

$$\frac{1}{2\gamma} - \frac{L}{2} - \frac{\gamma\eta}{2} \geq 0$$

and hence

$$\mathbb{E}\left[\Phi_{t+1}\right] \leq \mathbb{E}\left[\Phi_t\right] - \frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + 4\gamma c\delta(1+s)\zeta^2.$$

Therefore, summing up the above inequality for $t = 0, 1, \ldots, T-1$ and rearranging the terms, we get

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \leq \frac{1}{\kappa}\left(\frac{2}{\gamma T}\left(\mathbb{E}\left[\Phi_0\right] - \mathbb{E}\left[\Phi_T\right]\right) + 8c\delta(1+s)\zeta^2\right)$$
$$\leq \frac{1}{\kappa}\left(\frac{2}{\gamma T}\left(\mathbb{E}\left[\Phi_0\right]\right) + 8c\delta(1+s)\zeta^2\right)$$
$$= \frac{1}{\kappa}\left(\frac{2\Phi_0}{\gamma T} + 8c\delta(1+s)\zeta^2\right).$$

Letting $s = \sqrt{\frac{1}{8c\delta G}}$ gives the result. □

Corollary F.1. *Let the assumptions of Theorem 2.2 hold, $p = b/m$ and $\gamma = \left(L + \sqrt{\eta}\right)^{-1}$, where $\eta = \left(8\omega(2\omega+1)\left(L_\pm^2 + L^2\right) + \frac{1}{b}\left(12\omega\left(2\omega+1\right) + \frac{2}{p}\right)\mathcal{L}_\pm^2\right)\left(\sqrt{\frac{1}{G}} + \sqrt{8c\delta}\right)^2$. Then for all $T \geq 0$, $\mathbb{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right]$ is of the order*

$$\mathcal{O}\left(\left(\frac{\left(L + \sqrt{\omega(\omega+1)\left(L_\pm^2 + L^2\right) + \left(\frac{\omega(\omega+1)}{b} + \frac{m}{b^2}\right)\mathcal{L}_\pm^2}\left(\sqrt{\frac{1}{G}} + \sqrt{c\delta}\right)\right)\delta^0}{T\left(1 - \left(c\delta + \sqrt{\frac{c\delta}{G}}\right)B\right)} + \frac{\left(c\delta + \sqrt{\frac{c\delta}{G}}\right)\zeta^2}{1 - \left(c\delta + \sqrt{\frac{c\delta}{G}}\right)B}\right)\right),$$

where $\delta^0 = f(x^0) - f^*$ and $\widehat{x}^T$ is chosen uniformly at random from $x^0, x^1, \ldots, x^{T-1}$. Hence, to guarantee $\mathbb{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right] \leq \varepsilon^2$ for $\varepsilon^2 > \left(1 - \left(8c\delta + \sqrt{8c\delta/G}\right)B\right)^{-1}\left(8c\delta + \sqrt{8c\delta/G}\right)\zeta^2$, Byz-DASHA-PAGE *requires*

$$\mathcal{O}\left(\frac{1}{\varepsilon^2}\left(1 + \left(\omega + \frac{\sqrt{m}}{b}\right)\left(\sqrt{\frac{1}{G}} + \sqrt{c\delta}\right)\right)\right)$$

communication rounds.

G MISSING PROOFS FOR Byz-EF21 AND Byz-EF21-BC

For the sake of clarity, we again adopt the notation $R^t \overset{\text{def}}{=} \left\|x^{t+1}-x^t\right\|^2$, $H_i^t \overset{\text{def}}{=} \|g_i^t - \nabla f_i(x^t)\|^2$, $H^t \overset{\text{def}}{=} \frac{1}{G}\sum_{i\in\mathcal{G}} H_i^t$, and additionally denote $G^t \overset{\text{def}}{=} \|x^t - w^t\|^2$.

G.1 Technical Lemmas

Lemma G.1 (Descent Lemma). *Suppose that Assumptions 1.1 and 1.4 hold. Then*

$$\begin{aligned}\mathbb{E}\left[f(x^{t+1})\right] \;\leq\; & \mathbb{E}\left[f(x^t)\right] - \frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] - \left(\frac{1}{2\gamma}-\frac{L}{2}\right)\mathbb{E}\left[R^t\right] \\ & + \frac{\gamma}{2}\left(1+\sqrt{8c\delta}\right)^2\mathbb{E}\left[H^t\right] + 4\gamma c\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2,\end{aligned}$$

where $\kappa = 1 - 8Bc\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)$.

Proof. First, from Lemma E.2, for any $s>0$ we have

$$\begin{aligned}\mathbb{E}\left[f(x^{t+1})\right] \;\leq\; & \mathbb{E}\left[f(x^t)\right] - \frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] - \left(\frac{1}{2\gamma}-\frac{L}{2}\right)\mathbb{E}\left[R^t\right] + 4\gamma c\delta(1+s)\mathbb{E}\left[H^t\right] \\ & + \frac{\gamma}{2}(1+s^{-1})\mathbb{E}\left[\left\|\overline{g}^t - \nabla f(x^t)\right\|^2\right] + 4\gamma c\delta(1+s)\zeta^2.\end{aligned}$$

Moreover, since by Jensen's inequality

$$\begin{aligned}\mathbb{E}\left[\left\|\overline{g}^t - \nabla f(x^t)\right\|^2\right] \;=\; & \mathbb{E}\left[\left\|\frac{1}{G}\sum_{i\in\mathcal{G}}\left(g_i^t - \nabla f_i(x^t)\right)\right\|^2\right] \\ \overset{(11)}{\leq}\; & \frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^t - \nabla f_i(x^t)\right\|^2\right] = \mathbb{E}\left[H^t\right],\end{aligned}$$

we get

$$\begin{aligned}\mathbb{E}\left[f(x^{t+1})\right] \;\leq\; & \mathbb{E}\left[f(x^t)\right] - \frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] - \left(\frac{1}{2\gamma}-\frac{L}{2}\right)\mathbb{E}\left[R^t\right] \\ & + \frac{\gamma}{2}\left(8(1+s)c\delta + 1 + s^{-1}\right)\mathbb{E}\left[H^t\right] + 4\gamma c\delta(1+s)\zeta^2.\end{aligned}$$

Choosing $s = \frac{1}{\sqrt{8c\delta}}$ to minimize $8(1+s)c\delta + 1 + s^{-1}$ gives the result. □

Lemma G.2. *Let Assumptions 1.1 and 1.2 hold. Then*

$$\mathbb{E}\left[H^{t+1}\right] \;\leq\; \left(1-\frac{\alpha_D}{4}\right)\mathbb{E}\left[H^t\right] + \frac{8}{\alpha_D}\left(L_\pm^2+L^2\right)\mathbb{E}\left[R^t\right] + \frac{10}{\alpha_D}\left(L_\pm^2+L^2\right)\mathbb{E}\left[G^{t+1}\right]. \tag{15}$$

Proof. First, the update rule of the algorithm implies that

$$\begin{aligned}\mathbb{E}\left[H^{t+1}\right] \;=\; & \frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^{t+1} - \nabla f_i(x^{t+1})\right\|^2\right] \\ =\; & \frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^t + \mathcal{C}_i^D(\nabla f_i(w^{t+1}) - g_i^t) - \nabla f_i(x^{t+1})\right\|^2\right] \\ \overset{(8)}{\leq}\; & \left(1+\frac{\alpha_D}{2}\right)\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\mathcal{C}_i^D(\nabla f_i(w^{t+1}) - g_i^t) - (\nabla f_i(w^{t+1}) - g_i^t)\right\|^2\right] \\ & + \left(1+\frac{2}{\alpha_D}\right)\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\nabla f_i(w^{t+1}) - \nabla f_i(x^{t+1})\right\|^2\right]\end{aligned}$$

$$
\begin{aligned}
&\overset{(1.2)}{\leq} \left(1+\frac{\alpha_D}{2}\right)(1-\alpha_D)\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^t-\nabla f_i(w^{t+1})\right\|^2\right] \\
&\quad+\left(1+\frac{2}{\alpha_D}\right)\left(L_\pm^2+L^2\right)\mathbb{E}\left[\left\|w^{t+1}-x^{t+1}\right\|^2\right] \\
&\overset{(9)}{\leq} \left(1-\frac{\alpha_D}{2}\right)\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^t-\nabla f_i(w^{t+1})\right\|^2\right]+\left(1+\frac{2}{\alpha_D}\right)\left(L_\pm^2+L^2\right)\mathbb{E}\left[G^{t+1}\right]. \qquad (16)
\end{aligned}
$$

The first term on the right hand side of (16) can be bounded as

$$
\begin{aligned}
&\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^t-\nabla f_i(w^{t+1})\right\|^2\right] \\
&\overset{(8)}{\leq}\left(1+\frac{\alpha_D}{4}\right)\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^t-\nabla f_i(x^t)\right\|^2\right] \\
&\quad+\left(1+\frac{4}{\alpha_D}\right)\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|\nabla f_i(x^t)-\nabla f_i(w^{t+1})\right\|^2\right] \\
&\overset{(1.2)}{\leq}\left(1+\frac{\alpha_D}{4}\right)\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^t-\nabla f_i(x^t)\right\|^2\right]+\left(1+\frac{4}{\alpha_D}\right)\left(L_\pm^2+L^2\right)\mathbb{E}\left[\left\|x^t-w^{t+1}\right\|^2\right] \\
&\overset{(8)}{\leq}\left(1+\frac{\alpha_D}{4}\right)\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^t-\nabla f_i(x^t)\right\|^2\right]+2\left(1+\frac{4}{\alpha_D}\right)\left(L_\pm^2+L^2\right)\mathbb{E}\left[\left\|x^t-x^{t+1}\right\|^2\right] \\
&\quad+2\left(1+\frac{4}{\alpha_D}\right)\left(L_\pm^2+L^2\right)\mathbb{E}\left[\left\|x^{t+1}-w^{t+1}\right\|^2\right] \\
&=\left(1+\frac{\alpha_D}{4}\right)\mathbb{E}\left[H^t\right]+2\left(1+\frac{4}{\alpha_D}\right)\left(L_\pm^2+L^2\right)\mathbb{E}\left[R^t\right] \\
&\quad+2\left(1+\frac{4}{\alpha_D}\right)\left(L_\pm^2+L^2\right)\mathbb{E}\left[G^{t+1}\right].
\end{aligned}
$$

Applying the bound above in (16), we obtain

$$
\begin{aligned}
\mathbb{E}\left[H^{t+1}\right] &\overset{(16)}{\leq} \left(1-\frac{\alpha_D}{2}\right)\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^t-\nabla f_i(w^{t+1})\right\|^2\right]+\left(1+\frac{2}{\alpha_D}\right)\left(L_\pm^2+L^2\right)\mathbb{E}\left[G^{t+1}\right] \\
&\leq \left(1-\frac{\alpha_D}{2}\right)\left(1+\frac{\alpha_D}{4}\right)\mathbb{E}\left[H^t\right]+2\left(1-\frac{\alpha_D}{2}\right)\left(1+\frac{4}{\alpha_D}\right)\left(L_\pm^2+L^2\right)\mathbb{E}\left[R^t\right] \\
&\quad+2\left(1-\frac{\alpha_D}{2}\right)\left(1+\frac{4}{\alpha_D}\right)\left(L_\pm^2+L^2\right)\mathbb{E}\left[G^{t+1}\right] \\
&\quad+\left(1+\frac{2}{\alpha_D}\right)\left(L_\pm^2+L^2\right)\mathbb{E}\left[G^{t+1}\right] \\
&\overset{(9)}{\leq} \left(1-\frac{\alpha_D}{4}\right)\mathbb{E}\left[H^t\right]+2\left(\frac{4}{\alpha_D}-\frac{\alpha_D}{2}-1\right)\left(L_\pm^2+L^2\right)\mathbb{E}\left[R^t\right] \\
&\quad+\left(2\left(\frac{4}{\alpha_D}-\frac{\alpha_D}{2}-1\right)+1+\frac{2}{\alpha_D}\right)\left(L_\pm^2+L^2\right)\mathbb{E}\left[G^{t+1}\right] \\
&\leq \left(1-\frac{\alpha_D}{4}\right)\mathbb{E}\left[H^t\right]+\frac{8}{\alpha_D}\left(L_\pm^2+L^2\right)\mathbb{E}\left[R^t\right]+\frac{10}{\alpha_D}\left(L_\pm^2+L^2\right)\mathbb{E}\left[G^{t+1}\right]
\end{aligned}
$$

as required. □

Lemma G.3. *Let $\mathcal{C}^P$ be a contractive compressor. Then*

$$
\mathbb{E}\left[G^{t+1}\right]\leq\left(1-\frac{\alpha_P}{2}\right)\mathbb{E}\left[G^t\right]+\frac{2}{\alpha_P}\mathbb{E}\left[R^t\right]. \qquad (17)
$$

Proof. Using the update rule of w^t, we obtain

$$\begin{aligned}
\mathbb{E}\left[G^{t+1}\right] &= \mathbb{E}\left[\left\|w^{t+1}-x^{t+1}\right\|^2\right] \\
&= \mathbb{E}\left[\left\|w^t+\mathcal{C}^P\left(x^{t+1}-w^t\right)-x^{t+1}\right\|^2\right] \\
&\leq (1-\alpha_P)\mathbb{E}\left[\left\|x^{t+1}-w^t\right\|^2\right] \\
&\overset{(8)}{\leq} (1-\alpha_P)\left(1+\frac{\alpha_P}{2}\right)\mathbb{E}\left[\left\|x^t-w^t\right\|^2\right]+(1-\alpha_P)\left(1+\frac{2}{\alpha_P}\right)\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right] \\
&\overset{(9),(10)}{\leq} \left(1-\frac{\alpha_P}{2}\right)\mathbb{E}\left[\left\|x^t-w^t\right\|^2\right]+\frac{2}{\alpha_P}\mathbb{E}\left[\left\|x^{t+1}-x^t\right\|^2\right].
\end{aligned}$$

□

G.2 Proof of Theorem 3.1

We can now prove the main result.

Theorem 3.1. *Let Assumptions 1.1, 1.2, and 1.4 hold. Assume that $\mathcal{C}_i^D \in \mathbb{B}(\alpha_D)$, $\mathcal{C}^P \in \mathbb{B}(\alpha_P)$, $0<\gamma\leq(L+\sqrt{\eta})^{-1}$ and $\delta<(8c(\sqrt{B}+B)^2)^{-1}$, where $\eta=\frac{32}{\alpha_D^2}\left(1+\frac{5}{\alpha_P^2}\right)\left(1+\sqrt{8c\delta}\right)^2\left(L_\pm^2+L^2\right)$. Initialize $w^0=x^0$, and $g_i^0=\nabla f_i(x^0)$ for all $i\in\mathcal{G}$. Then for all $T\geq 0$, the iterates produced by* Byz-EF21/Byz-EF21-BC *satisfy*

$$\mathbb{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right]\leq\frac{1}{A}\left(\frac{2\delta^0}{\gamma T}+\left(8c\delta+\sqrt{8c\delta}\right)\zeta^2\right),$$

where $\delta^0=f(x^0)-f^*$, $A=1-B\left(8c\delta+\sqrt{8c\delta}\right)$ and $\widehat{x}^T$ is chosen uniformly at random from $x^0,x^1,\ldots,x^{T-1}$.

Proof. We start with Lemma G.1:

$$\begin{aligned}
\mathbb{E}\left[f(x^{t+1})\right] \leq\ & \mathbb{E}\left[f(x^t)\right]-\frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right]-\left(\frac{1}{2\gamma}-\frac{L}{2}\right)\mathbb{E}\left[R^t\right] \\
&+\frac{\gamma}{2}\left(1+\sqrt{8c\delta}\right)^2\mathbb{E}\left[H^t\right]+4\gamma c\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2.
\end{aligned}$$

Adding a $\frac{2\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2$ multiple of (15) gives

$$\begin{aligned}
\mathbb{E}\left[f(x^{t+1})\right]+\frac{2\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2\mathbb{E}\left[H^{t+1}\right] &\leq \mathbb{E}\left[f(x^t)\right]-\frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \\
&-\left(\frac{1}{2\gamma}-\frac{L}{2}\right)\mathbb{E}\left[R^t\right]+\frac{\gamma}{2}\left(1+\sqrt{8c\delta}\right)^2\mathbb{E}\left[H^t\right]+4\gamma c\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2 \\
&+\frac{2\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2\left(1-\frac{\alpha_D}{4}\right)\mathbb{E}\left[H^t\right] \\
&+\underbrace{\frac{2\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2\frac{8}{\alpha_D}\left(L_\pm^2+L^2\right)}_{\frac{\gamma\nu}{2}}\mathbb{E}\left[R^t\right] \\
&+\underbrace{\frac{2\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2\frac{10}{\alpha_D}\left(L_\pm^2+L^2\right)}_{\frac{5\gamma\nu}{8}}\mathbb{E}\left[G^{t+1}\right] \\
&= \mathbb{E}\left[f(x^t)\right]+\frac{2\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2\mathbb{E}\left[H^t\right]-\frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \\
&-\left(\frac{1}{2\gamma}-\frac{L}{2}-\frac{\gamma\nu}{2}\right)\mathbb{E}\left[R^t\right]+4\gamma c\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2+\frac{5\gamma\nu}{8}\mathbb{E}\left[G^{t+1}\right],
\end{aligned}$$

where $\nu \overset{\text{def}}{=} \frac{32}{\alpha_D^2}\left(1+\sqrt{8c\delta}\right)^2\left(L_\pm^2+L^2\right)$. Next, adding a $\frac{5\gamma\nu}{4\alpha_P}$ multiple of (17), we obtain

$$\begin{aligned}
&\mathbb{E}\left[f(x^{t+1})\right]+\frac{2\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2\mathbb{E}\left[H^{t+1}\right]+\frac{5\gamma\nu}{4\alpha_P}\mathbb{E}\left[G^{t+1}\right]\\
&\qquad\leq \mathbb{E}\left[f(x^t)\right]+\frac{2\gamma}{\alpha}\left(1+\sqrt{8c\delta}\right)^2\mathbb{E}\left[H^t\right]-\frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right]\\
&\qquad\quad-\left(\frac{1}{2\gamma}-\frac{L}{2}-\frac{\gamma\nu}{2}\right)\mathbb{E}\left[R^t\right]+4\gamma c\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2\\
&\qquad\quad+\frac{5\gamma\nu}{8}\mathbb{E}\left[G^{t+1}\right]+\frac{5\gamma\nu}{4\alpha_P}\left(1-\frac{\alpha_P}{2}\right)\mathbb{E}\left[G^t\right]+\frac{5\gamma\nu}{4\alpha_P}\frac{2}{\alpha_P}\mathbb{E}\left[R^t\right]\\
&\qquad= \mathbb{E}\left[f(x^t)\right]+\frac{2\gamma}{\alpha}\left(1+\sqrt{8c\delta}\right)^2\mathbb{E}\left[H^t\right]-\frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right]\\
&\qquad\quad-\left(\frac{1}{2\gamma}-\frac{L}{2}-\frac{\gamma\eta}{2}\right)\mathbb{E}\left[R^t\right]+4\gamma c\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2+\frac{5\gamma\nu}{8}\mathbb{E}\left[G^{t+1}\right]\\
&\qquad\quad+\frac{5\gamma\nu}{4\alpha_P}\left(1-\frac{\alpha_P}{2}\right)\mathbb{E}\left[G^t\right],
\end{aligned}$$

where $\eta \overset{\text{def}}{=} \nu+\frac{5\nu}{\alpha_P^2}$. Rearranging the above inequality, we get

$$\begin{aligned}
&\mathbb{E}\left[f(x^{t+1})\right]+\frac{2\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2\mathbb{E}\left[H^{t+1}\right]+\frac{5\gamma\nu}{4\alpha_P}\left(1-\frac{\alpha_P}{2}\right)\mathbb{E}\left[G^{t+1}\right]\\
&\qquad\leq \mathbb{E}\left[f(x^t)\right]+\frac{2\gamma}{\alpha}\left(1+\sqrt{8c\delta}\right)^2\mathbb{E}\left[H^t\right]-\frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right]\\
&\qquad\quad-\left(\frac{1}{2\gamma}-\frac{L}{2}-\frac{\gamma\eta}{2}\right)\mathbb{E}\left[R^t\right]+\frac{5\gamma\nu}{4\alpha_P}\left(1-\frac{\alpha_P}{2}\right)\mathbb{E}\left[G^t\right]+4\gamma c\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2\\
&\qquad\leq \mathbb{E}\left[f(x^t)\right]+\frac{2\gamma}{\alpha}\left(1+\sqrt{8c\delta}\right)^2\mathbb{E}\left[H^t\right]-\frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right]\\
&\qquad\quad+\frac{5\gamma\nu}{4\alpha_P}\left(1-\frac{\alpha_P}{2}\right)\mathbb{E}\left[G^t\right]+4\gamma c\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2,
\end{aligned}$$

where the last inequality follows from the fact that by Lemma C.2 and our assumption on the stepsize we have $\frac{1}{2\gamma}-\frac{L}{2}-\frac{\gamma\eta}{2}>0$. Hence, for $\Psi^t \overset{\text{def}}{=} f(x^t)-f^*+\frac{2\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2 H^t+\frac{5\gamma\nu}{4\alpha_P}\left(1-\frac{\alpha_P}{2}\right)G^t$, we obtain

$$\mathbb{E}\left[\Psi^{t+1}\right]\leq \mathbb{E}\left[\Psi^t\right]-\frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right]+4\gamma c\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2,$$

Summing the terms for $t=0,\ldots,T-1$ gives

$$\mathbb{E}\left[\Psi^T\right]\leq \mathbb{E}\left[\Psi^0\right]-\frac{\gamma\kappa}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right]+4\gamma c\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)T\zeta^2.$$

Since from our assumption on δ it follows that $\kappa>0$, rearranging the terms and dividing by T, we obtain

$$\begin{aligned}
\sum_{t=0}^{T-1}\frac{1}{T}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right]&\leq \frac{2}{\gamma\kappa T}\mathbb{E}\left[\Psi^0\right]-\frac{2}{\gamma\kappa T}\mathbb{E}\left[\Psi^T\right]+\frac{8c\delta}{\kappa}\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2\\
&\leq \frac{2\Psi^0}{\gamma\kappa T}+\frac{8c\delta}{\kappa}\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2,
\end{aligned}$$

which proves the result. □

Corollary G.1. *Let the assumptions of Theorem 3.1 hold and* $\gamma=\left(L+\sqrt{\eta}\right)^{-1}$, *where* $\eta=\frac{32}{\alpha_D^2}\left(1+\frac{5}{\alpha_P^2}\right)\left(1+\sqrt{8c\delta}\right)^2\left(L_\pm^2+L^2\right)$. *Then for all* $T\geq 0$, $\mathbb{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right]$ *with* $\widehat{x}^T$ *chosen uniformly at random*

from $x^0, x^1, \ldots, x^{T-1}$ *is of order*

$$\mathcal{O}\left(\frac{\left(L + \frac{1+\sqrt{c\delta}}{\alpha_D}\sqrt{\left(1+\frac{1}{\alpha_P^2}\right)\left(L_\pm^2 + L^2\right)}\right)\delta^0}{T\left(1-\left(c\delta+\sqrt{c\delta}\right)B\right)} + \frac{\left(c\delta+\sqrt{c\delta}\right)\zeta^2}{1-\left(c\delta+\sqrt{c\delta}\right)B}\right),$$

where $\delta^0 = f(x^0) - f^*$.

Hence, to guarantee $\mathbb{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right] \leq \varepsilon^2$ *for* $\varepsilon^2 > 8c\delta\left(1 - 8Bc\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)\right)^{-1}\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2$, Byz-EF21/Byz-EF21-BC *requires*

$$\mathcal{O}\left(\frac{1+\sqrt{c\delta}}{\alpha_D\alpha_P\varepsilon^2}\right)$$

communication rounds.

H CONVERGENCE FOR POLYAK-ŁOJASIEWICZ FUNCTIONS

The $\mathcal{O}(1/T)$ rate obtained for general smooth non-convex objective functions can be improved to a fast linear rate without assuming strong convexity of f upon employing a weaker Polyak-Łojasiewicz condition.

Assumption H.1 (Polyak-Łojasiewicz condition)**.** *The function f satisfies Polyak-Łojasiewicz (PŁ) condition with parameter μ, i.e., for all $x \in \mathbb{R}^d$ there exists $x^* \in \arg\min_{x\in\mathbb{R}^d} f(x)$ such that*

$$2\mu\left(f(x) - f(x^*)\right) \leq \left\|\nabla f(x)\right\|^2. \tag{18}$$

Below we prove the convergence results of our methods under the above assumption.

H.1 Byz-VR-MARINA 2.0

Lemma H.1 (Descent lemma)**.** *Suppose that Assumptions 1.1, 1.4 and H.1 hold. Then for all $s > 0$ we have*

$$\begin{aligned}\mathbb{E}\left[f(x^{t+1}) - f^*\right] \quad \leq \quad & (1 - \gamma\mu\kappa)\,\mathbb{E}\left[f(x^t) - f^*\right] - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\mathbb{E}\left[R^t\right] + 4\gamma c\delta(1+s)\mathbb{E}\left[H^t\right] \\ & + \frac{\gamma}{2}(1 + s^{-1})\mathbb{E}\left[\left\|\overline{g}^t - \nabla f(x^t)\right\|^2\right] + 4\gamma c\delta(1+s)\zeta^2,\end{aligned}$$

where $\kappa = 1 - 8Bc\delta(1+s)$.

Proof. The result follows from combining Lemma E.2 with Assumption H.1. □

Theorem H.1. *Let Assumptions 1.1, 1.2, 1.3, 1.4, and H.1 hold. Suppose $g_i^0 = \nabla f_i(x^0)$ for all $i \in \mathcal{G}$ and*

$$0 < \gamma \leq \min\left\{\frac{1}{L + \sqrt{\eta}}, \frac{p}{2\mu}\right\}, \qquad 0 < \delta < \frac{1}{(8c + 4\sqrt{c})\,B},$$

where $\eta = 2\frac{1-p}{p}\left(\omega\left(\frac{\mathcal{L}_\pm^2}{b} + L_\pm^2 + L^2\right) + \frac{\mathcal{L}_\pm^2}{b}\right)\left(\sqrt{\frac{1}{G}} + \sqrt{8c\delta}\right)^2$. Then for all $T \geq 0$ the iterates produced by Byz-VR-MARINA 2.0 *satisfy*

$$\begin{aligned}\mathbb{E}\left[f(x^T) - f^*\right] \quad \leq \quad & \left(1 - \gamma\mu\left(1 - \left(8c\delta + \sqrt{\frac{8c\delta}{G}}\right)B\right)\right)^T \mathbb{E}\left[f(x^0) - f^*\right] \\ & + \frac{\left(4c\delta + \sqrt{\frac{2c\delta}{G}}\right)\zeta^2}{\mu\left(1 - \left(8c\delta + \sqrt{\frac{8c\delta}{G}}\right)B\right)}.\end{aligned}$$

where $\Phi_0 = f(x^0) - f^* + \gamma\left(\frac{1+s^{-1}}{p}\left\|\overline{g}^0 - \nabla f\left(x^0\right)\right\|^2 + \frac{8c\delta(1+s)}{p}H^t\right)$.

Proof. Let $C = \frac{2}{p}(1 + s^{-1})$ and $D = \frac{16}{p}c\delta\,(1+s)$ for some $s > 0$ and denote

$$\begin{aligned}\psi_t &= C\left\|\overline{g}^t - \nabla f(x^t)\right\|^2 + DH^t, \\ \Phi_t &= f(x^t) - f^* + \frac{\gamma}{2}\psi_t.\end{aligned}$$

Using lemmas H.1, E.3 and E.4, we have:

$$\begin{aligned}\mathbb{E}\left[\Phi_{t+1}\right] \quad &= \quad \mathbb{E}\left[f(x^{t+1}) - f^* + \frac{\gamma}{2}\psi_{t+1}\right] \\ &= \quad \underbrace{\mathbb{E}\left[f(x^{t+1}) - f^*\right]}_{\text{use H.1}} + \frac{\gamma}{2}\left(C\underbrace{\mathbb{E}\left[\left\|\overline{g}^{t+1} - \nabla f(x^{t+1})\right\|^2\right]}_{\text{use E.3}} + D\underbrace{\mathbb{E}\left[H^{t+1}\right]}_{\text{use E.4}}\right)\end{aligned}$$

$$
\begin{aligned}
&\leq (1-\gamma\mu\kappa)\,\mathbb{E}\left[f(x^t)-f^*\right] \\
&\quad +\frac{\gamma}{2}(1-\gamma\mu\kappa)\Bigg(\underbrace{C\mathbb{E}\left[\left\|\overline{g}^t-\nabla f\left(x^t\right)\right\|^2\right]+D\mathbb{E}\left[H^t\right]}_{\mathbb{E}[\psi_t]}\Bigg) \\
&\quad -\left(\frac{1}{2\gamma}-\frac{L}{2}-\frac{\gamma\eta}{2}\right)\mathbb{E}\left[R^t\right]+4\gamma c\delta(1+s)\zeta^2 \\
&= (1-\gamma\mu\kappa)\,\mathbb{E}\left[\Phi_t\right]-\left(\frac{1}{2\gamma}-\frac{L}{2}-\frac{\gamma\eta}{2}\right)\mathbb{E}\left[R^t\right]+4\gamma c\delta(1+s)\zeta^2,
\end{aligned}
$$

where

$$
\begin{aligned}
\eta &\overset{\text{def}}{=} C(1-p)\left(\frac{\omega}{G}\left(\frac{\mathcal{L}_\pm^2}{b}+L_\pm^2+L^2\right)+\frac{\mathcal{L}_\pm^2}{Gb}\right)+D(1-p)\left(\omega\left(\frac{\mathcal{L}_\pm^2}{b}+L_\pm^2+L^2\right)+\frac{\mathcal{L}_\pm^2}{b}\right) \\
&= 2\frac{1-p}{p}\left(\omega\left(\frac{\mathcal{L}_\pm^2}{b}+L_\pm^2+L^2\right)+\frac{\mathcal{L}_\pm^2}{b}\right)\left(\frac{1+s^{-1}}{G}+8c\delta(1+s)\right).
\end{aligned}
$$

Taking $0<\gamma\leq\frac{1}{L+\sqrt{\eta}}$ and using Lemma C.2 gives

$$
\frac{1}{2\gamma}-\frac{L}{2}-\frac{\gamma\eta}{2}\geq 0,
$$

and hence

$$
\mathbb{E}\left[\Phi_{t+1}\right]\leq(1-\gamma\mu\kappa)\,\mathbb{E}\left[\Phi_t\right]+4\gamma c\delta(1+s)\zeta^2.
$$

Similarly as in the proof of Theorem 2.1, taking $s=\frac{1}{\sqrt{8c\delta G}}$, unrolling the recurrence and rearranging the terms, we get

$$
\begin{aligned}
\mathbb{E}\left[\Phi_T\right] &\leq (1-\gamma\mu\kappa)^T\,\mathbb{E}\left[\Phi_0\right]+4c\delta\gamma(1+s)\zeta^2\sum_{t=0}^{T}(1-\gamma\mu\kappa)^t \\
&\leq (1-\gamma\mu\kappa)^T\,\mathbb{E}\left[\Phi_0\right]+4c\delta\gamma(1+s)\zeta^2\sum_{t=0}^{\infty}(1-\gamma\mu\kappa)^t \\
&= (1-\gamma\mu\kappa)^T\,\mathbb{E}\left[\Phi_0\right]+\frac{4c\delta(1+s)\zeta^2}{\mu\kappa} \\
&= \left(1-\gamma\mu\left(1-\left(8c\delta+\sqrt{\frac{8c\delta}{G}}\right)B\right)\right)^T\mathbb{E}\left[f(x^0)-f^*\right] \\
&\quad +\frac{\left(4c\delta+\sqrt{\frac{2c\delta}{G}}\right)\zeta^2}{\mu\left(1-\left(8c\delta+\sqrt{\frac{8c\delta}{G}}\right)B\right)}.
\end{aligned}
$$

The result follows from the fact that $\Phi_T\geq f(x^T)-f^*$. □

H.2 Byz-DASHA-PAGE

Lemma H.2 (Descent lemma). *Suppose that Assumptions 1.1, 1.4 and H.1 hold. Then for all $s>0$ we have*

$$
\begin{aligned}
\mathbb{E}\left[f(x^{t+1})-f^*\right] &\leq (1-\gamma\mu\kappa)\,\mathbb{E}\left[f(x^t)-f^*\right]-\left(\frac{1}{2\gamma}-\frac{L}{2}\right)\mathbb{E}\left[R^t\right] \\
&\quad +8\gamma c\delta(1+s)\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^t-h_i^t\right\|^2\right]\right)+8\gamma c\delta(1+s)\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]\right) \\
&\quad +\gamma(1+s^{-1})\mathbb{E}\left[\left\|\overline{g}^t-h^t\right\|^2\right]+\gamma(1+s^{-1})\mathbb{E}\left[\left\|h^t-\nabla f(x^t)\right\|^2\right]+4\gamma c\delta(1+s)\zeta^2,
\end{aligned}
$$

where $\kappa=1-8Bc\delta(1+s)$.

Proof. The result follows from combining Lemma F.1 with Assumption H.1. □

Lemma H.3 (Calculations). *Let* $\kappa = 1 - 8Bc\delta(1+s)$ *and assume that*

$$0 < a \leq \frac{1}{2\omega+1},$$
$$0 < \gamma < \min\left\{\frac{p}{2\mu\kappa}, \frac{a}{2\mu\kappa}\right\}.$$

The inequalities

$$\begin{cases} C, D, E, F \geq 0 \\ (1-\gamma\mu\kappa)\, C \geq (1-a)^2 C + 2\left(1+s^{-1}\right) \\ (1-\gamma\mu\kappa)\, D \geq \left(2a^2\omega + (1-a)^2\right) D + \frac{2a^2\omega}{G} C + 16c\delta\,(1+s) \\ (1-\gamma\mu\kappa)\, E \geq (1-p)\, E + 2\left(1+s^{-1}\right) \\ (1-\gamma\mu\kappa)\, F \geq (1-p)\, F + 4p\omega\left(\frac{C}{G} + D\right) + 16c\delta\,(1+s) \end{cases} \tag{19}$$

are satisfied when

$$\begin{cases} C = \frac{2(1+s^{-1})}{1-(1-a)^2-\frac{a}{2}} \\ D = \frac{1}{1-2\omega a^2-(1-a)^2-\frac{a}{2}}\left(16\,(1+s)\,c\delta + \frac{2\omega a^2}{G} C\right) \\ E = \frac{4(1+s^{-1})}{p} \\ F = 8\omega\left(\frac{C}{G} + D\right) + \frac{32c\delta}{p}\,(1+s)\,. \end{cases}$$

Further, for this choice of C, D, E, F, *we have*

$$\begin{aligned} \frac{C}{G} + D &= \frac{2}{1-2\omega a^2-(1-a)^2-\frac{a}{2}}\left(\frac{1+s^{-1}}{G} + (1+s)\,8c\delta\right) \\ \frac{E}{G} + F &= \left(\frac{16\omega}{1-2\omega a^2-(1-a)^2-\frac{a}{2}} + \frac{4}{p}\right)\left(\frac{1+s^{-1}}{G} + (1+s)\,8c\delta\right). \end{aligned} \tag{20}$$

Proof. Let $w = \gamma\mu\kappa$. Under the assumption that $0 < \gamma < \min\{\frac{p}{2\mu\kappa}, \frac{a}{2\mu\kappa}\}$, we have that $0 < w < \min\{\frac{p}{2}, \frac{a}{2}\}$. Furthermore, (19) implies that

$$\begin{cases} C \geq 2\frac{1+s^{-1}}{1-(1-a)^2-w} \\ D \geq \frac{1}{1-2\omega a^2-(1-a)^2-w}\left(16\,(1+s)\,c\delta + \frac{2\omega a^2}{G} C\right) \\ E \geq 2\frac{1+s^{-1}}{p-w} \\ F \geq \frac{4\omega p}{p-w}\left(\frac{C}{G} + D\right) + \frac{16c\delta}{p-w}\,(1+s) \end{cases}$$

The assumption that $0 < a \leq \frac{1}{2\omega+1}$ ensures that $1 - 2\omega a^2 - (1-a)^2 - \frac{a}{2} > 0$. Since $0 < w < \min\{\frac{p}{2}, \frac{a}{2}\}$, one can take

$$\begin{cases} C = 2\frac{1+s^{-1}}{1-(1-a)^2-\frac{a}{2}} \\ D = \frac{1}{1-2\omega a^2-(1-a)^2-\frac{a}{2}}\left(16\,(1+s)\,c\delta + \frac{2\omega a^2}{G} C\right) \\ E = 4\frac{1+s^{-1}}{p} \\ F = 8\omega\left(\frac{C}{G} + D\right) + \frac{32c\delta}{p}\,(1+s)\,. \end{cases}$$

It remains to verify (20) by direct substitution of the values above. □

Theorem H.2 (General Convergence). *Let Assumptions 1.1, 1.2, 1.4, 1.3, and H.1 hold. Let*

$$\begin{cases} C = \frac{2(1+\sqrt{8c\delta G})}{1-(1-a)^2-\frac{a}{2}} \\ D = \frac{1}{1-2\omega a^2-(1-a)^2-\frac{a}{2}}\left(16\left(1+\frac{1}{\sqrt{8c\delta G}}\right)c\delta + \frac{2\omega a^2}{G} C\right) \\ E = \frac{4(1+\sqrt{8c\delta G})}{p} \\ F = 8\omega\left(\frac{C}{G} + D\right) + \frac{32c\delta}{p}\left(1+\frac{1}{\sqrt{8c\delta G}}\right) \end{cases}$$

and assume that

$$0 < \gamma \leq \min\left\{\frac{1}{L+\sqrt{\eta}}, \frac{p}{2\mu\kappa}, \frac{a}{2\mu\kappa}\right\}, \qquad 0 < a \leq \frac{1}{2\omega+1}, \qquad 0 < \delta < \frac{1}{8c\left(1+\frac{1}{\sqrt{8c\delta G}}\right)B},$$

where $\eta = \left(\frac{8\omega\left(L_{\pm}^2+L^2\right)}{1-2\omega a^2-(1-a)^2-\frac{a}{2}} + \frac{1-p}{b}\mathcal{L}_{\pm}^2\left(\frac{20\omega}{1-2\omega a^2-(1-a)^2-\frac{a}{2}}+\frac{4}{p}\right)\right)\left(\frac{1+\sqrt{8c\delta G}}{G}+\left(1+\frac{1}{\sqrt{8c\delta G}}\right)8c\delta\right)$ *and* $\kappa = 1 - 8Bc\delta\left(1+\frac{1}{\sqrt{8c\delta G}}\right)$. *Then for all* $T \geq 0$ *the iterates produced by* Byz-DASHA-PAGE *satisfy*

$$\mathbb{E}\left[f(x^T)-f^*\right] \leq \left(1-\gamma\mu\left(1-\left(8c\delta+\sqrt{\frac{8c\delta}{G}}\right)B\right)\right)^T \mathbb{E}\left[f(x^0)-f^*\right] + \frac{\left(4c\delta+\sqrt{\frac{2c\delta}{G}}\right)\zeta^2}{\mu\left(1-\left(8c\delta+\sqrt{\frac{8c\delta}{G}}\right)B\right)}.$$

Proof. Let

$$\begin{aligned}
\psi_t &= C\left\|\overline{g}_t - h^t\right\|^2 + D\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\left\|g_i^t - h_i^t\right\|^2\right) + E\left\|h^t - \nabla f(x^t)\right\|^2 \\
&\quad + F\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\left\|h_i^t - \nabla f_i(x^t)\right\|^2\right), \\
\Phi_t &= f(x^t) - f^* + \frac{\gamma}{2}\psi_t.
\end{aligned}$$

Using lemmas H.2, F.3, F.4, F.5 and F.6, we have

$$\begin{aligned}
\mathbb{E}\left[\Phi_{t+1}\right] &= \mathbb{E}\left[f(x^{t+1}) - f^* + \frac{\gamma}{2}\psi_{t+1}\right] \\
&= \underbrace{\mathbb{E}\left[f(x^{t+1})-f^*\right]}_{\text{use H.2}} + \frac{\gamma}{2}C\underbrace{\mathbb{E}\left[\left\|\overline{g}^{t+1}-h^{t+1}\right\|^2\right]}_{\text{use F.3}} + \frac{\gamma}{2}D\underbrace{\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^{t+1}-h_i^{t+1}\right\|^2\right]\right)}_{\text{use F.4}} \\
&\quad + \frac{\gamma}{2}E\underbrace{\mathbb{E}\left[\left\|h^{t+1}-\nabla f(x^{t+1})\right\|^2\right]}_{\text{use F.5}} + \frac{\gamma}{2}F\underbrace{\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^{t+1}-\nabla f_i(x^{t+1})\right\|^2\right]\right)}_{\text{use F.6}} \\
&\leq (1-\gamma\mu\kappa)\,\mathbb{E}\left[f(x^t)-f^*\right] \\
&\quad + (1-\gamma\mu\kappa)\frac{\gamma}{2}\left(C\mathbb{E}\left[\left\|\overline{g}_t-h^t\right\|^2\right] + D\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|g_i^t-h_i^t\right\|^2\right]\right)\right. \\
&\quad \left. + E\mathbb{E}\left[\left\|h^t-\nabla f(x^t)\right\|^2\right] + F\left(\frac{1}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\left\|h_i^t-\nabla f_i(x^t)\right\|^2\right]\right)\right) \\
&\quad - \left(\frac{1}{2\gamma}-\frac{L}{2}-\frac{\gamma\eta}{2}\right)\mathbb{E}\left[R^t\right] + 4\gamma c\delta(1+s)\zeta^2 \\
&= (1-\gamma\mu\kappa)\,\mathbb{E}\left[\Phi_t\right] - \left(\frac{1}{2\gamma}-\frac{L}{2}-\frac{\gamma\eta}{2}\right)\mathbb{E}\left[R^t\right] + 4\gamma c\delta(1+s)\zeta^2,
\end{aligned}$$

where the inequality holds by Lemma H.3 and

$$\eta = 2\omega\left(\frac{1-p}{b}\mathcal{L}_{\pm}^2 + 2\left(L_{\pm}^2+L^2\right)\right)\left(\frac{C}{G}+D\right) + \frac{1-p}{B}\mathcal{L}_{\pm}^2\left(\frac{E}{G}+F\right)$$

$$= \left(\frac{8\omega\left(L_{\pm}^2 + L^2\right)}{1 - 2\omega a^2 - (1-a)^2 - \frac{a}{2}} + \frac{1-p}{b}\mathcal{L}_{\pm}^2 \left(\frac{20\omega}{1 - 2\omega a^2 - (1-a)^2 - \frac{a}{2}} + \frac{4}{p} \right) \right) \times \left(\frac{1 + s^{-1}}{G} + (1+s)\, 8c\delta \right).$$

Using the assumption on the stepsize and Lemma C.2, we have

$$\frac{1}{2\gamma} - \frac{L}{2} - \frac{\gamma\eta}{2} \geq 0$$

and hence

$$\mathbb{E}\left[\Phi_{t+1}\right] \leq (1 - \gamma\mu\kappa)\, \mathbb{E}\left[\Phi_t\right] + 4\gamma c\delta(1+s)\zeta^2.$$

Unrolling the recurrence and rearranging the terms, we get

$$\begin{aligned}
\mathbb{E}\left[\Phi_T\right] &\leq (1 - \gamma\mu\kappa)^T\, \mathbb{E}\left[\Phi_0\right] + 4c\delta\gamma(1+s)\zeta^2 \sum_{t=0}^{T} (1 - \gamma\mu\kappa)^t \\
&\leq (1 - \gamma\mu\kappa)^T\, \mathbb{E}\left[\Phi_0\right] + 4c\delta\gamma(1+s)\zeta^2 \sum_{t=0}^{\infty} (1 - \gamma\mu\kappa)^t \\
&= (1 - \gamma\mu\kappa)^T\, \mathbb{E}\left[\Phi_0\right] + \frac{4c\delta(1+s)\zeta^2}{\mu\kappa}.
\end{aligned}$$

Noting that $\Phi_T \geq f(x^T) - f^*$ and letting $s = \frac{1}{\sqrt{8c\delta G}}$ gives the result. □

H.3 Byz-EF21 and Byz-EF21-BC

Theorem H.3. *Let Assumptions 1.1, 1.2, 1.4 and H.1 hold and suppose that*

$$0 < \gamma \leq \min\left\{ \frac{1}{L + \sqrt{\eta}}, \frac{\alpha_D}{8\kappa\mu}, \frac{\alpha_P}{4\kappa\mu} \right\}, \qquad \delta < \frac{1}{8c(\sqrt{B} + B)^2}, \tag{21}$$

where $\eta = \frac{64}{\alpha_D^2}\left(1 + \frac{10}{\alpha_P^2}\left(1 - \frac{\alpha_P}{4}\right)\right)\left(1 + \sqrt{8c\delta}\right)^2\left(L_{\pm}^2 + L^2\right)$ *and* $\kappa = 1 - 8Bc\delta\left(1 + \frac{1}{\sqrt{8c\delta}}\right)$. *Then for all* $T \geq 0$ *the iterates of* Byz-EF21-BC *satisfy*

$$\mathbb{E}\left[f(x^T) - f^*\right] \quad \leq \quad \left(1 - \gamma\mu\left(1 - \left(8c\delta + \sqrt{8c\delta}\right)B\right)\right)^T \Psi^0 + \frac{\left(4c\delta + \sqrt{2c\delta}\right)\zeta^2}{\left(1 - \left(8c\delta + \sqrt{8c\delta}\right)B\right)\mu},$$

where $\Psi^0 \overset{\text{def}}{=} f(x^t) - f^* + \frac{4\gamma}{\alpha_D}\left(1 + \sqrt{8c\delta}\right)^2 H^0 + \frac{320\gamma}{2\alpha_P\alpha_D^2}\left(1 - \frac{\alpha_P}{2}\right)\left(1 + \sqrt{8c\delta}\right)^2\left(L_{\pm}^2 + L^2\right)G^0$.

Proof. Adding a $\frac{4\gamma}{\alpha_D}\left(1 + \sqrt{8c\delta}\right)^2$ multiple of (15) to the inequality from Lemma G.1 gives

$$\begin{aligned}
&\mathbb{E}\left[f(x^{t+1})\right] + \frac{4\gamma}{\alpha_D}\left(1 + \sqrt{8c\delta}\right)^2 \mathbb{E}\left[H^{t+1}\right] \\
&\quad\leq \mathbb{E}\left[f(x^t)\right] - \frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\mathbb{E}\left[R^t\right] \\
&\qquad + \frac{\gamma}{2}\left(1 + \sqrt{8c\delta}\right)^2 \mathbb{E}\left[H^t\right] + 4\gamma c\delta\left(1 + \frac{1}{\sqrt{8c\delta}}\right)\zeta^2 \\
&\qquad + \frac{4\gamma}{\alpha_D}\left(1 + \sqrt{8c\delta}\right)^2\left(1 - \frac{\alpha_D}{4}\right)\mathbb{E}\left[H^t\right] \\
&\qquad + \underbrace{\frac{4\gamma}{\alpha_D}\left(1 + \sqrt{8c\delta}\right)^2 \frac{8}{\alpha_D}\left(L_{\pm}^2 + L^2\right)}_{\frac{\gamma\nu}{2}} \mathbb{E}\left[R^t\right]
\end{aligned}$$

$$
\begin{aligned}
&+ \underbrace{\frac{4\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2 \frac{10}{\alpha_D}\left(L_{\pm}^2+L^2\right)}_{\frac{5\gamma\nu}{8}} \mathbb{E}\left[G^{t+1}\right] \\
&= \mathbb{E}\left[f(x^t)\right] + \frac{4\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2\left(1-\frac{\alpha_D}{8}\right)\mathbb{E}\left[H^t\right] - \frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \\
&\quad - \left(\frac{1}{2\gamma}-\frac{L}{2}-\frac{\gamma\nu}{2}\right)\mathbb{E}\left[R^t\right] + 4\gamma c\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2 + \frac{5\gamma\nu}{8}\mathbb{E}\left[G^{t+1}\right],
\end{aligned}
$$

where $\nu \overset{\text{def}}{=} \frac{64}{\alpha_D^2}\left(1+\sqrt{8c\delta}\right)^2\left(L_{\pm}^2+L^2\right)$. Adding $\frac{5\gamma\nu}{2\alpha_P}\left(1-\frac{\alpha_P}{4}\right)$ multiple of (17), we obtain

$$
\begin{aligned}
&\mathbb{E}\left[f(x^{t+1})\right] + \frac{4\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2\mathbb{E}\left[H^{t+1}\right] + \frac{5\gamma\nu}{2\alpha_P}\left(1-\frac{\alpha_P}{4}\right)\mathbb{E}\left[G^{t+1}\right] \\
&\le \mathbb{E}\left[f(x^t)\right] + \frac{4\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2\left(1-\frac{\alpha_D}{8}\right)\mathbb{E}\left[H^t\right] - \frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \\
&\quad - \left(\frac{1}{2\gamma}-\frac{L}{2}-\frac{\gamma\nu}{2}\right)\mathbb{E}\left[R^t\right] + 4\gamma c\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2 + \frac{5\gamma\nu}{8}\mathbb{E}\left[G^{t+1}\right] \\
&\quad + \frac{5\gamma\nu}{2\alpha_P}\left(1-\frac{\alpha_P}{4}\right)\left(1-\frac{\alpha_P}{2}\right)\mathbb{E}\left[G^t\right] + \frac{5\gamma\nu}{2\alpha_P}\left(1-\frac{\alpha_P}{4}\right)\frac{2}{\alpha_P}\mathbb{E}\left[R^t\right] \\
&= \mathbb{E}\left[f(x^t)\right] + \frac{4\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2\left(1-\frac{\alpha_D}{8}\right)\mathbb{E}\left[H^t\right] - \frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \\
&\quad - \left(\frac{1}{2\gamma}-\frac{L}{2}-\frac{\gamma\eta}{2}\right)\mathbb{E}\left[R^t\right] + 4\gamma c\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2 + \frac{5\gamma\nu}{8}\mathbb{E}\left[G^{t+1}\right] \\
&\quad + \frac{5\gamma\nu}{2\alpha_P}\left(1-\frac{\alpha_P}{4}\right)\left(1-\frac{\alpha_P}{2}\right)\mathbb{E}\left[G^t\right],
\end{aligned}
$$

where $\eta \overset{\text{def}}{=} \nu + \frac{10\nu}{\alpha_P^2}\left(1-\frac{\alpha_P}{4}\right)$. Rearranging the above inequality gives

$$
\begin{aligned}
&\mathbb{E}\left[f(x^{t+1})\right] + \frac{4\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2\mathbb{E}\left[H^{t+1}\right] + \frac{5\gamma\nu}{2\alpha_P}\left(1-\frac{\alpha_P}{2}\right)\mathbb{E}\left[G^{t+1}\right] \\
&\le \mathbb{E}\left[f(x^t)\right] + \frac{4\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2\left(1-\frac{\alpha_D}{8}\right)\mathbb{E}\left[H^t\right] - \frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \\
&\quad - \left(\frac{1}{2\gamma}-\frac{L}{2}-\frac{\gamma\eta}{2}\right)\mathbb{E}\left[R^t\right] + 4\gamma c\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2 \\
&\quad + \frac{5\gamma\nu}{2\alpha_P}\left(1-\frac{\alpha_P}{2}\right)\left(1-\frac{\alpha_P}{4}\right)\mathbb{E}\left[G^t\right] \\
&\le \mathbb{E}\left[f(x^t)\right] + \frac{4\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2\left(1-\frac{\alpha_D}{8}\right)\mathbb{E}\left[H^t\right] - \frac{\gamma\kappa}{2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \\
&\quad + 4\gamma c\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2 + \frac{5\gamma\nu}{2\alpha_P}\left(1-\frac{\alpha_P}{2}\right)\left(1-\frac{\alpha_P}{4}\right)\mathbb{E}\left[G^t\right],
\end{aligned}
$$

where in the last inequality we use the fact that $\frac{1}{2\gamma}-\frac{L}{2}-\frac{\gamma\eta}{2} \ge 0$ by Lemma C.2 and our assumption on the stepsize. Now, let us define $\Psi^t = f(x^t) - f^* + \frac{4\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2 H^t + \frac{5\gamma\nu}{2\alpha_P}\left(1-\frac{\alpha_P}{2}\right)G^t$. Subtracting f^* from both sides of the above inequality and noting that our assumption on δ implies that $\kappa > 0$, the PŁ inequality (Assumption H.1) gives

$$
\begin{aligned}
\mathbb{E}\left[\Psi^{t+1}\right] &= \mathbb{E}\left[f(x^{t+1})-f^*\right] + \frac{4\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2\mathbb{E}\left[H^{t+1}\right] + \frac{5\gamma\nu}{2\alpha_P}\left(1-\frac{\alpha_P}{2}\right)\mathbb{E}\left[G^{t+1}\right] \\
&\overset{\text{(H.1)}}{\le} \mathbb{E}\left[f(x^t)-f^*\right] + \frac{4\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2\left(1-\frac{\alpha_D}{8}\right)\mathbb{E}\left[H^t\right] - \gamma\kappa\mu\mathbb{E}\left[f(x^t)-f^*\right] \\
&\quad + 4\gamma c\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2 + \frac{5\gamma\nu}{2\alpha_P}\left(1-\frac{\alpha_P}{2}\right)\left(1-\frac{\alpha_P}{4}\right)\mathbb{E}\left[G^t\right]
\end{aligned}
$$

$$
\begin{aligned}
&= (1-\gamma\kappa\mu)\,\mathbb{E}\left[f(x^t)-f^*\right] + \frac{4\gamma}{\alpha_D}\left(1+\sqrt{8c\delta}\right)^2\left(1-\frac{\alpha_D}{8}\right)\mathbb{E}\left[H^t\right] \\
&\quad + \frac{5\gamma\nu}{2\alpha_P}\left(1-\frac{\alpha_P}{2}\right)\left(1-\frac{\alpha_P}{4}\right)\mathbb{E}\left[G^t\right] + 4\gamma c\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2 \\
&\overset{(21)}{\leq} (1-\gamma\kappa\mu)\,\mathbb{E}\left[\Psi^t\right] + 4\gamma c\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2,
\end{aligned}
$$

where in the last step we use the assumption on γ to establish that $1-\frac{\alpha_D}{8} \leq 1-\gamma\kappa\mu$ and $1-\frac{\alpha_P}{4} \leq 1-\gamma\kappa\mu$. Applying the inequality iteratively gives

$$
\begin{aligned}
\mathbb{E}\left[\Psi^T\right] &\leq (1-\gamma\kappa\mu)^T\,\mathbb{E}\left[\Psi^0\right] + 4\gamma c\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2\sum_{t=0}^{T-1}(1-\gamma\kappa\mu)^t \\
&\leq (1-\gamma\kappa\mu)^T\,\mathbb{E}\left[\Psi^0\right] + 4\gamma c\delta\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2\sum_{t=0}^{\infty}(1-\gamma\kappa\mu)^t \\
&= (1-\gamma\kappa\mu)^T\,\mathbb{E}\left[\Psi^0\right] + \frac{4c\delta}{\kappa\mu}\left(1+\frac{1}{\sqrt{8c\delta}}\right)\zeta^2.
\end{aligned}
$$

Noting that $\mathbb{E}\left[\Psi^T\right] \geq \mathbb{E}\left[f(x^T)-f^*\right]$, we finish the proof. □

I NUMERICAL EXPERIMENTS: ADDITIONAL DETAILS AND RESULTS

I.1 Logistic regression experiments

Objective: We consider solving the following logistic regression problem with regularization r:

$$\min_{x\in\mathbb{R}^d}\left\{f(x)=\frac{1}{N}\sum_{i=1}^{N}\log\left(1+\exp\left(-y_i a_i^T x\right)\right)+\frac{\lambda}{2}r(x)\right\},$$

where d is the dimension of the model, N is the total number of samples, a_i^T is the ith row of the design matrix $A\in\mathbb{R}^{N\times d}$, $y_i\in\{-1,1\}$ are the corresponding labels and $\lambda>0$ is the regularization parameter. We choose $\lambda=0.1$ in all experiments. The regularizer is chosen depending on the nature of the problem we are treating:

- **Non-convex case:** a non-convex regularizer $r:x\mapsto\sum_{j=1}^{d}\frac{x_j^2}{1+x_j^2}$
- **PŁ case:** the ridge regularization $r:x\mapsto\|x\|^2$. It can be shown that for this chice of r, for $\lambda>0$, the objective function is strongly convex and therefore satisfies the PŁ condition.

Datasets: We consider two `LibSVM` datasets (Chang and Lin, 2011) (summarised in Table 2). In each case, the data is distributed among $n=16$ workers, out of which 3 are Byzantine.

Table 2: Overview of the `LibSVM` datasets used.

Dataset	N (# samples)	d (# features)
`phishing`	11,055	68
`w8a`	49,749	300

As for the data distribution among the workers, we consider two scenarios:

- **Homogeneous setting:** all the workers (regular and Byzantine) have access to the entire dataset (hence the gradients calculated by the good workers are all equal).
- **Heterogeneous setting:** the data is evenly distributed among good workers, without any uniform sampling or shuffling, so that each worker has access to approximately the same amount of data and there is no overlap. Since we make no assumptions on the behaviour of the bad workers, they have access to the full dataset.

Byzantine attacks and aggregation rule: In each setting, we consider 4 different byzantine attacks:

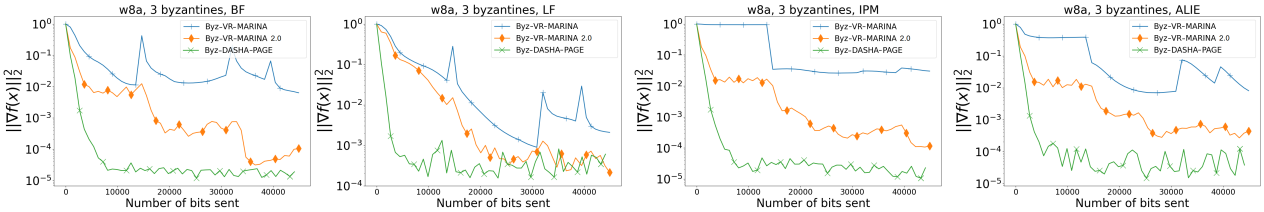
- **Bit Flipping (BF)**: Byzantine workers compute $-\nabla f(x)$ and send it to the server.
- **Label Flipping (LF)**: Byzantine workers compute their gradients using poisoned labels (i.e., $y_i\to -y_i$)
- **A Little Is Enough (ALIE)** (Baruch et al., 2019): Byzantine workers compute empirical mean μ_G and standard deviation σ_G of $\{g_i^t\}_{i\in G}$ and send $\mu_G-z\sigma_G$ to the server, where z is a constant that controls the strength of the attack.
- **Inner Product Manipulation (IPM)** (Xie et al., 2020): Byzantine workers send $-\frac{z}{G}\sum_{i\in\mathcal{G}}\nabla f_i(x)$ to the server, where $z>0$ is a constant that controls the strength of the attack.

The aggregation rule is the Coordinate-wise Median (CM) aggregator (Yin et al., 2018) with bucketing (Karimireddy et al., 2022) (see Appendix B).

Parameters choice: For each method, we finetune the stepsize within the set $\{2^k \times \gamma_{th}\}_{k\in\mathbb{N}}$, where γ_{th} is the theoretical stepsize of the algorithm under consideration (indicated by $1\times, 2\times, 4\times, \ldots$ in the plots). The momentum parameter a and probability p are chosen to be the optimal values predicted by theory, and the value of the constant c from Definition 1.1 is determined using the formula from Karimireddy et al. (2022).

In algorithms with stochastic gradients, uniform sampling with no replacement and batch size $b = 0.01m$ is used. In experiments comparing methods using unbiased compressors, we use the Rand$K$ compressor with $K = 0.1d$. In error feedback experiments (Appendix I.1.3), Byz-VR-MARINA and Byz-EF21 employ the Rand$K$ and Top$K$ sparsifiers, respectively, with $K = 1$.

I.1.1 Extra dataset

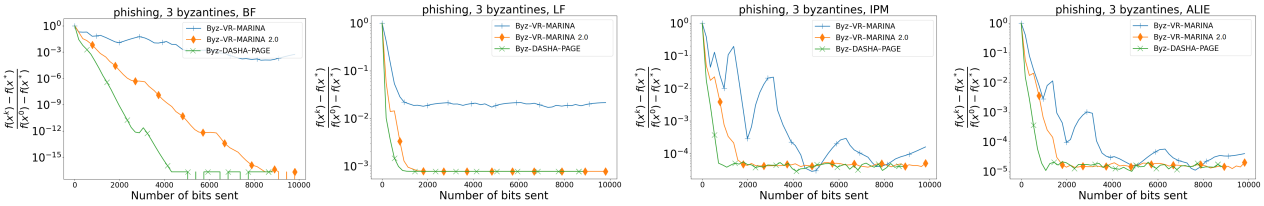


(a) BF attack (b) LF attack (c) IPM attack (d) ALIE attack

Figure 3: Communication complexity comparison in the heterogeneous non-convex setting on the `w8a` dataset.

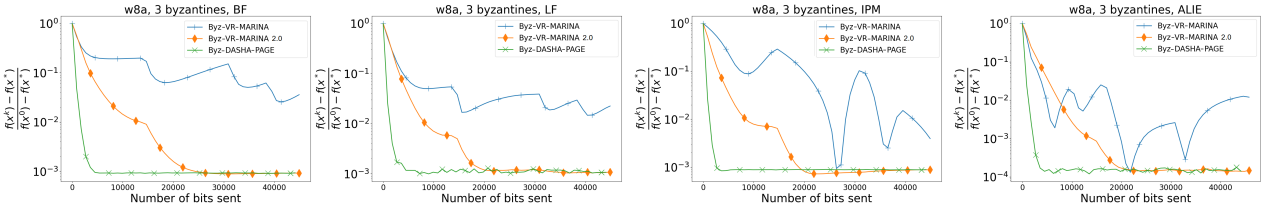
Similarly to the results obtained with the `phishing` dataset, both Byz-VR-MARINA 2.0 and Byz-DASHA-PAGE have superior performance compared to Byz-VR-MARINA when tested on the `w8a` dataset (Figure 3), across all types of attacks. Our methods converge faster and achieve better accuracy.

I.1.2 Convergence under Polyak-Łojasiewicz condition



(a) BF attack (b) LF attack (c) IPM attack (d) ALIE attack

Figure 4: Communication complexity comparison in the heterogeneous strongly convex setting on the `phishing` dataset.



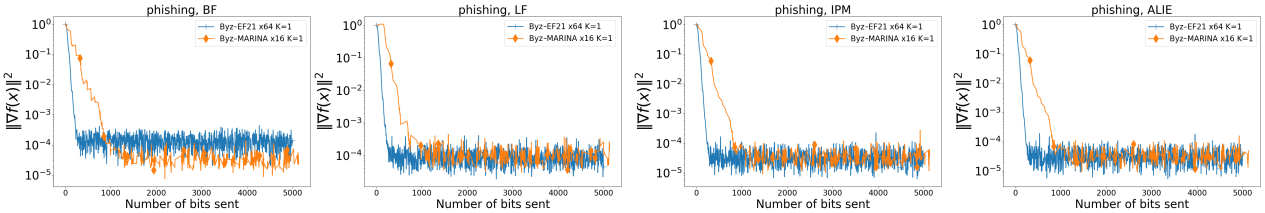
(a) BF attack (b) LF attack (c) IPM attack (d) ALIE attack

Figure 5: Communication complexity comparison in the heterogeneous strongly convex setting on the `w8a` dataset.

As indicated by Theorems H.1 and H.2, the experiments in the heterogeneous Polyak-Łojasiewicz setting (Figures 4 and 5) confirm that our methods outperform Byz-VR-MARINA, converging faster, being more stable and achieving higher accuracy. The improvement is most pronounced in the case of BF and LF attacks.

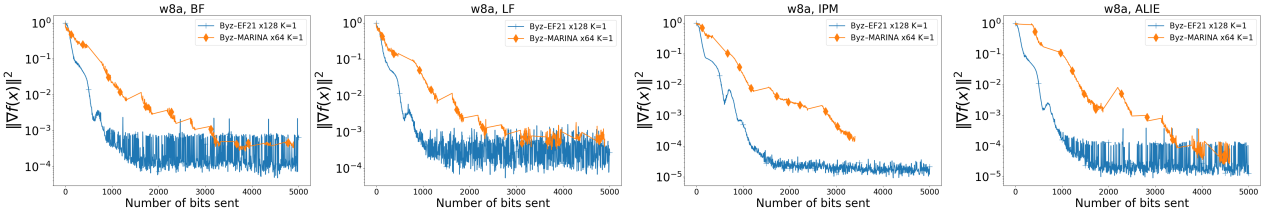
I.1.3 Error Feedback experiments

We next compare the empirical performance of Byz-EF21 and Byz-VR-MARINA in the heterogeneous non-convex setting. To ensure fair comparison, full gradients are calculated.



(a) BF attack (b) LF attack (c) IPM attack (d) ALIE attack

Figure 6: Communication complexity comparison in the heterogeneous non-convex setting on the `phishing` dataset.



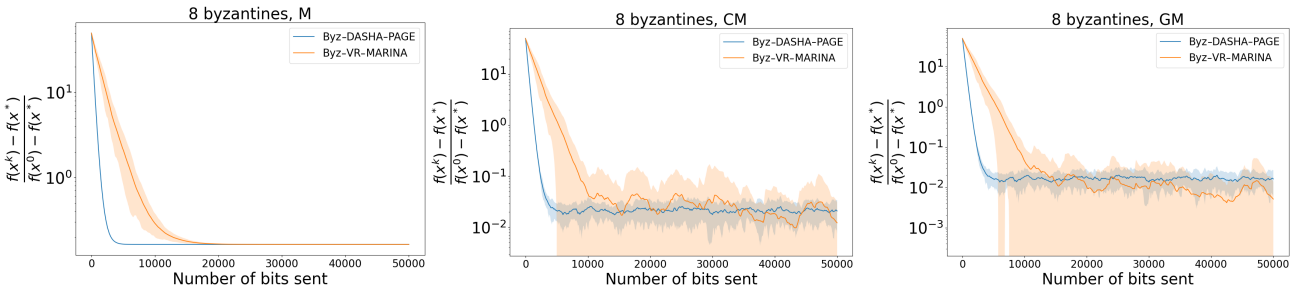
(a) BF attack (b) LF attack (c) IPM attack (d) ALIE attack

Figure 7: Communication complexity comparison in the heterogeneous non-convex setting on the `w8a` dataset.

The experiments unveil promising potential for communication improvement. As shown in Figures 6 and 7, Byz-EF21 converges faster than Byz-VR-MARINA, before both reach a point of stagnation.

I.2 Neighborhood size

We next compare the accuracy of Byz-DASHA-PAGE and Byz-VR-MARINA. The data is distributed among 28 clients, out of which 8 are Byzantine. The 20 good clients are divided into two groups, $\mathcal{G}_1$ and $\mathcal{G}_2$, of equal size. The local functions calculated by the clients are $f_i(x) = \frac{1}{2}\|x\|^2 + \langle \zeta_1, x\rangle$ for $i \in \mathcal{G}_1$ and $f_i(x) = \frac{1}{2}\|x\|^2 + \langle \zeta_2, x\rangle$ for $i \in \mathcal{G}_2$ where $\zeta_1, \zeta_2 \in \mathbb{R}^d$ with $d = 100$. The Byzantines mimic the behavior of group $\mathcal{G}_1$. 3 different aggregation rules are considered: standard averaging, CM aggregator and GM aggregator (see Appendix B). In both algorithms, we use the Rand$K$ compressor with $K = 5$ and stepsize equal to 10 times the theoretical one. The results, presented in Figure 8, not only show that Byz-DASHA-PAGE converges faster, but it is also more stable, converging to a smaller neighborhood than Byz-VR-MARINA.



(a) Mean aggregator (b) CM aggregator (c) GM aggregator

Figure 8: The mean optimality gap of Byz-DASHA-PAGE and Byz-VR-MARINA for 3 aggregation rules calculated from 15 runs of both algorithms. The shadowed area corresponds to one standard deviation.

I.3 Numerical Comparison of Theoretical Stepsizes

Table 3 gives the comparison of theoretical stepsizes calculated for the `phishing` dataset.

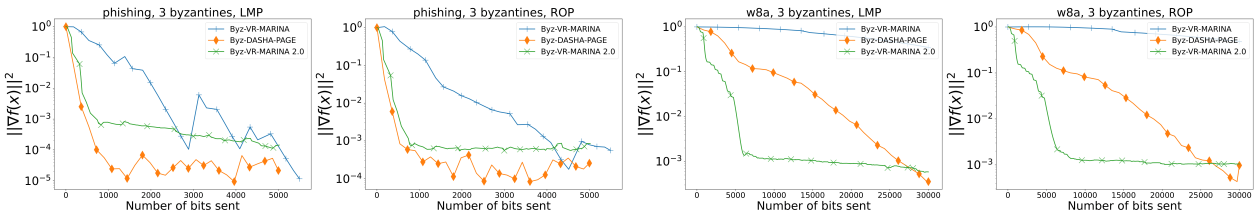
Table 3: Comparison of theoretical stepsizes for `phishing` dataset.

Byz-VR-MARINA	Byz-VR-MARINA 2.0	Byz-DASHA-PAGE
4e-4	2e-2	1.2e-2

The theory behind Byz-VR-MARINA 2.0 and Byz-DASHA-PAGE allows for significantly larger stepsize than Byz-VR-MARINA, as a result of which our methods require fewer communication rounds. This improvement also shows in experiments, where Byz-VR-MARINA 2.0 and Byz-DASHA-PAGE tolerate much larger stepsizes, which makes them more efficient in practice.

I.4 Convergence under Additional Attacks

In this subsection, we compare the convergence of Byz-VR-MARINA, Byz-VR-MARINA 2.0, and Byz-DASHA-PAGE on the non-convex logistic regression problem under two additional attacks: Local Model Poisoning (LMP) attack (Fang et al., 2020) and Relocated Orthogonal Perturbation (ROP) attack (Özfatura et al., 2023) adjusted to the considered methods. The results are given in Figure 9. Similarly to the results under previously considered attacks, Byz-VR-MARINA 2.0 and Byz-DASHA-PAGE outperform Byz-VR-MARINA under LMP and ROP attacks.

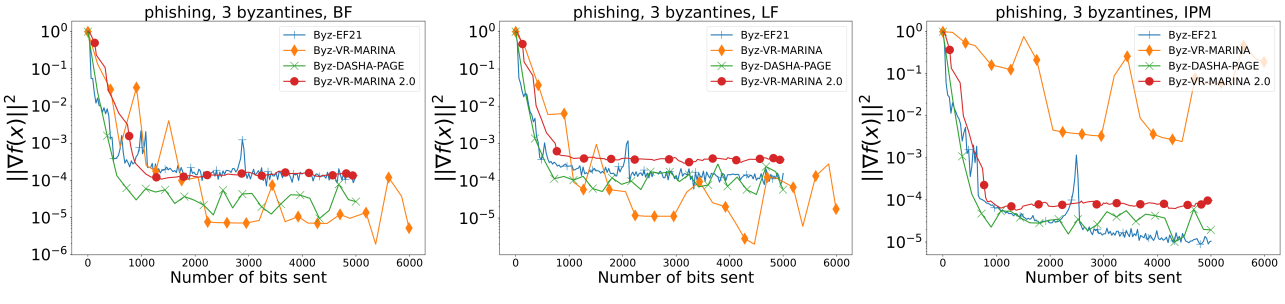


(a) LMP attack, `phishing` (b) ROP attack, `phishing` (c) LMP attack, `phishing` (d) ROP attack, `phishing`

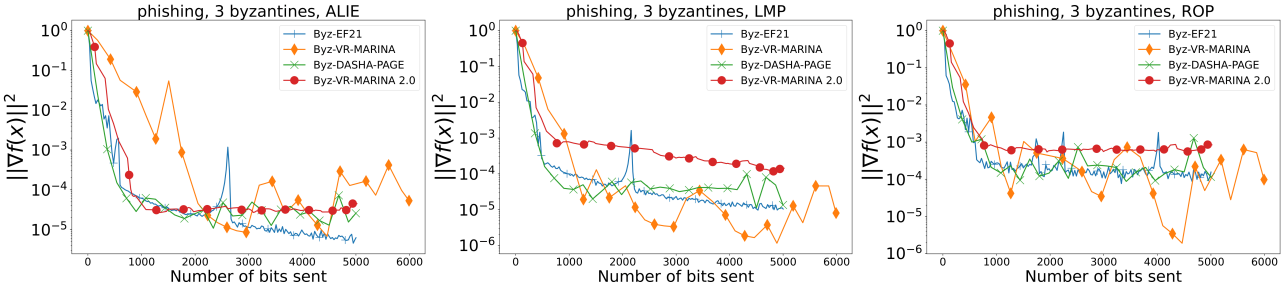
Figure 9: Communication complexity comparison in the heterogeneous non-convex setting on the `phishing` and `w8a` datasets under LMP and ROP attacks.

I.5 Biased Compression vs Unbiased Compression

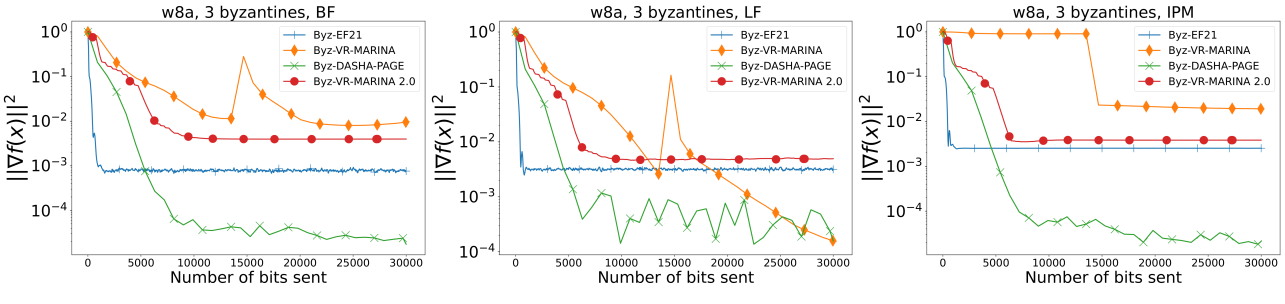
In this subsection, we compare Byz-EF21 and Byz-VR-MARINA, Byz-VR-MARINA 2.0, Byz-DASHA-PAGE with full gradient computations, i.e., we provide the comparison of the behavior of the methods with biased and unbiased compression. As in the previous subsection, we consider logistic regression problem with non-convex regularization. The results are given in Figure 10. In general, we see that the new methods are more robust than Byz-VR-MARINA. However, in the conducted experiments, there is no clear "champion", e.g., under IPM, LMP and ROP attacks for `phishing` dataset Byz-EF21 outperforms Byz-DASHA-PAGE and Byz-VR-MARINA 2.0, but Byz-DASHA-PAGE works noticeably better than Byz-EF21 under all attacks for `w8a` dataset.



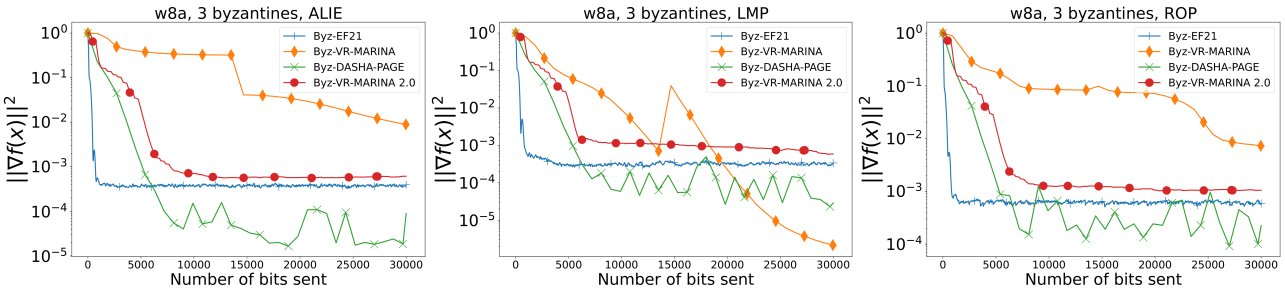
(a) BF attack, **phishing** (b) LF attack, **phishing** (c) IPM attack, **phishing**



(d) ALIE attack, **phishing** (e) LMP attack, **phishing** (f) ROP attack, **phishing**



(g) BF attack, **phishing** (h) LF attack, **phishing** (i) IPM attack, **phishing**



(j) ALIE attack, **phishing** (k) LMP attack, **phishing** (l) ROP attack, **phishing**

Figure 10: Communication complexity comparison in the heterogeneous non-convex setting on the **phishing** and **w8a** datasets for the methods with biased and unbiased compression.