Accelerating Byzantine-Robust Distributed Learning with Compressed Communication via Double Momentum and Variance Reduction

Yanghao Li Sichuan University
Changxin Liu† East China University of Science and Technology
Yuhao Yi† Sichuan University

Abstract

In collaborative and distributed learning, Byzantine robustness reflects a major facet of optimization algorithms. Such distributed algorithms are often accompanied by transmitting a large number of parameters, so communication compression is essential for an effective solution. In this paper, we propose Byz-DM21, a novel Byzantine-robust and communication-efficient stochastic distributed learning algorithm. Our key innovation is a novel gradient estimator based on a double-momentum mechanism, integrating recent advancements in error feedback techniques. Using this estimator, we design both standard and accelerated algorithms that eliminate the need for large batch sizes while maintaining robustness against Byzantine workers. We prove that the Byz-DM21 algorithm has a smaller neighborhood size and converges to ε-stationary points in $\mathcal{O}(\varepsilon^{-4})$ iterations. To further enhance efficiency, we introduce a distributed variant called Byz-VR-DM21, which incorporates local variance reduction at each node to progressively eliminate variance from random approximations. We show that Byz-VR-DM21 provably converges to ε-stationary points in $\mathcal{O}(\varepsilon^{-3})$ iterations. Additionally, we extend our results to the case where the functions satisfy the Polyak-Łojasiewicz condition. Finally, numerical experiments demonstrate the effectiveness of the proposed method.

1 INTRODUCTION

Distributed learning has garnered significant attention due to its widespread applications. Specific applications include large-scale training of deep neural networks, collaborative learning in edge computing, and distributed optimization in federated learning Haddadpour et al. (2019); Jaggi et al. (2014); Lee et al. (2017); Yu et al. (2019). Traditional distributed learning often assumes an environment free from faults or attacks. However, in certain real-world scenarios, such as edge computing Shi et al. (2016) and federated learning McMahan and Ramage (2017), service providers (also known as the server) typically have limited control over computational nodes (also known as workers). In such cases, workers may experience various software and hardware failures Xie et al. (2019). Worse still, some workers may be compromised by malicious third parties and intentionally send erroneous information to disrupt the distributed learning process Kairouz et al. (2021). Workers affected by such failures or attacks are referred to as *Byzantine workers*[1]. Distributed learning in the presence of Byzantine workers, also known as Byzantine-Robust Distributed Learning (BRDL), has recently emerged as a prominent research topic Yin et al. (2018); Bernstein et al. (2019); Diakonikolas and Kane (2019); Konstantinidis and Ramamoorthy (2021).

A common approach to achieving Byzantine robustness is to replace the standard mean aggregator with robust alternatives, such as Krum Blanchard et al. (2017), geometric median Chen et al. (2017), coordinate-wise median Yin et al. (2018), and trimmed mean Yin et al. (2018), among others. However, in the presence of Byzantine workers, even robust aggregators inevitably introduce aggregation error, defined as the discrepancy between the aggregated result and the true mean. Moreover, even with independent and identically distributed (i.i.d.) data, the aggregation error can be significant due to the high variance of stochastic gra-

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. †Corresponsing author.

[1]Following the standard terminology in the literature Lamport et al. (2019); Su and Vaidya (2016), we refer to a worker as Byzantine if it may, either maliciously or unintentionally, send incorrect information to other workers or to the server. Such workers are assumed to be omniscient: they can access the vectors transmitted by other workers, are aware of the server-side aggregation rule, and may coordinate their actions with one another.

dients Karimireddy et al. (2021), which are typically sent from workers to the server for parameter updates. Large aggregation errors may lead to the failure of BRDL methods Xie et al. (2020).

In addition to Byzantine robustness, the efficiency of distributed learning systems is considered a major performance metric. Due to its distributed nature, a bottleneck lies in the communication between workers and the server, particularly the transmission of local stochastic gradients. This challenge becomes more pronounced with high-dimensional models, resulting in substantial communication overhead. To mitigate this issue, several strategies have been proposed, including reducing communication frequency by performing multiple local updates Chen et al. (2018) and compressing the transmitted messages. Common compression techniques include quantization, which encodes vectors using a limited number of bits Alistarh et al. (2017), and sparsification, which reduces the number of non-zero elements in transmitted vectors Wangni et al. (2018). In this work, we primarily focus on the latter approach. Specifically, at each iteration, workers compress their local gradients before transmission, and the server aggregates these compressed gradients to update the model parameters.

Byzantine robustness and communication efficiency are both crucial properties in distributed learning, yet their simultaneous exploration has been relatively limited in the existing literature, with current methods facing notable challenges. Zhu and Ling (2021) proposed Byzantine-robust variants of compressed SGD (BR-CSGD) and SAGA (BR-CSAGA), as well as BROADCAST, which integrates DIANA Mishchenko et al. (2019) with BR-CSAGA. However, their convergence analysis is confined to strongly convex problems and relies on stringent assumptions. Similarly, Gorbunov et al. (2023) studied the Byzantine-tolerant Byz-VR-MARINA, which achieves fast convergence but occasionally triggers uncompressed message communication and full gradient computation. Moreover, most existing Byzantine-robust methods utilize unbiased compressors, whereas biased contractive compressors combined with error feedback often yield superior empirical performance Rammal et al. (2024).

In this work, we comprehensively address these limitations by building on the recently proposed Byzantine-robust stochastic distributed learning method with error feedback, Byz-EF21-SGDM Liu et al. (2026). We first propose Byz-DM21, a Byzantine-robust stochastic distributed learning algorithm that utilizes Double Momentum with Error Feedback-21, an enhanced variant of Byz-EF21-SGDM, featuring improved convergence properties over the original method. A double momentum estimator $u_i^{(t)}$ has richer "memory" of the past gradients compared to SGDM Fatkhullin et al. (2023). Building on Byz-DM21, we further introduce Byz-VR-DM21, which incorporates local variance reduction at each node to progressively eliminate variance arising from stochastic approximations. We summarize our main contributions as follows:

- We propose a novel Byzantine-robust and communication-efficient stochastic distributed learning method, Byz-DM21. Our new algorithm is batch-free and employs a Double-Momentum mechanism to simultaneously suppress variance from stochastic gradients and bias introduced by compression, which leads to improved sample complexity for Byz-EF21-SGDM in the non-asymptotic regime (see Remark 3.2). Moreover, we prove that the variance of the double-momentum estimator is strictly smaller than that of the single-momentum estimator, with an asymptotic reduction factor of approximately $1/2$ when η is small. We also show that Byz-DM21 converges to an ε-stationary point in $\mathcal{O}(\varepsilon^{-4})$ iterations.
- We propose Byz-VR-DM21 as an extension to Byz-DM21. The improved algorithm incorporates local variance reduction on all nodes to progressively eliminate variance from stochastic approximations. By combining worker momentum based variance reduction with a Byzantine robust aggregator, we obtain a faster Byzantine robust algorithm. We prove that Byz-VR-DM21 achieves an accelerated convergence to ε-stationary points in $\mathcal{O}(\varepsilon^{-3})$ iterations. Additionally, we analyze Byz-DM21 and Byz-VR-DM21 for problems that satisfy the Polyak-Łojasiewicz condition Polyak (1963).
- We derive complexity bounds for Byz-DM21 under standard assumptions and further extend these results to Byz-VR-DM21. These complexity bounds demonstrate that our algorithm outperforms the state-of-the-art, specifically in the full gradient setting Rammal et al. (2024), in terms of convergence speed. Notably, our results are tight and align with established lower bounds in both stochastic and full gradient scenarios when no Byzantine workers are present.
- Under the ζ^2-heterogeneity assumption, our algorithm converges to a tighter neighborhood around the optimal solution and also matches the established lower bound Karimireddy et al. (2021); Allouah et al. (2023). For a detailed comparison, see Table 1. Furthermore, experiments demonstrate that the proposed algorithm not only converges faster but also asymptotically reaches a model with a smaller error.

Table 1: Summary of related works on Byzantine-robust and communication-efficient distributed methods. "Complexity (NC)" and "Complexity (PŁ)": represents the total number of communication rounds required for each worker to find x such that $\mathbb{E}\left[\|\nabla f(x)\|\right] \leq \varepsilon$ in the general non-convex case and such x that $\mathbb{E}[f(x)-f(x^*)] \leq \varepsilon$ in PŁ case respectively. σ^2 represents the variance of local stochastic gradients, κ refers to the parameter of robust aggregators, $\alpha \in (0,1]$ and $\omega \geq 0$ are parameters for biased contractive and unbiased compressors, respectively, ζ^2 denotes the heterogeneity bound among honest workers, and c denotes the heterogeneity constant. The parameter $p \in (0,1]$ is the sampling probability used in Byz-VR-MARINA and Byz-DASHA-PAGE. m represents the local dataset size for workers in Byz-VR-MARINA,Byz-DASHA-PAGE and BROADCAST.

Method	Setting	Batch-free?	Complexity (NC)	Complexity (PŁ)	Accuracy
BROADCAST Zhu and Ling (2021)	finite-sum	✗	-	$\frac{m^2(1+\omega)^{3/2}G}{\mu^2(n-2B)}$ [(1)]	$\kappa(1+\omega)\zeta^2$
Byz-VR-MARINA [(2)] Gorbunov et al. (2023)	finite-sum	✗	$\frac{\left(1+\sqrt{\max\{\omega^2,m\omega\}}\left(\sqrt{\frac{1}{G}}+\sqrt{\kappa\max\{\omega,m\}}\right)\right)}{\varepsilon^2}$	$\frac{\left(1+\sqrt{\max\{\omega^2,m\omega\}}\left(\sqrt{\frac{1}{G}}+\sqrt{\kappa\max\{\omega,m\}}\right)\right)+\mu(m+\omega)}{\mu}$	$\frac{\kappa\zeta^2}{p-c\kappa}$
Byz-DASHA-PAGE [(2)] Rammal et al. (2024)	finite-sum	✗	$\frac{\left(1+\left(\omega+\frac{\sqrt{m}}{\omega}\right)\left(\sqrt{\frac{1}{G}}+\sqrt{\kappa}\right)\right)}{\varepsilon^2}$	-	$\frac{\kappa\zeta^2}{1-c\kappa}$
Byz-EF21 [(2)] Rammal et al. (2024)	full gradient	✗	$\frac{1+\sqrt{\kappa}}{\alpha\varepsilon^2}$	-	$\frac{(\kappa+\sqrt{\kappa})\zeta^2}{1-c(\kappa+\sqrt{\kappa})}$
Byz-EF21-SGDM Liu et al. (2026)	stochastic gradient	✔	$\frac{\sigma^2}{G\varepsilon^4}+\frac{\kappa\sigma^2}{\varepsilon^4}$	-	$\kappa\zeta^2$
Byz-DM21 (This work)	stochastic gradient	✔	$\frac{\sqrt{\kappa+1}\sigma^2}{G\varepsilon^4}+\frac{(\kappa+1)^{3/2}\sigma^2}{\varepsilon^4}$ $\frac{\sqrt{\kappa+1}}{\alpha\varepsilon^2}$ (full gradient)	$\frac{(G(\kappa+1)+1)\sigma^2(\mu+\sqrt{\kappa+1})}{\mu^2\varepsilon G}$	$\kappa\zeta^2$
Byz-VR-DM21 (This work)	stochastic gradient	✔	$\frac{\sqrt{\kappa+1}\sigma}{\sqrt{G}\varepsilon^3}+\frac{(\kappa+1)\sigma}{\varepsilon^3}$ $\frac{\sqrt{\kappa+1}}{\alpha\varepsilon^2}$ (full gradient)	$\frac{(G(\kappa+1)+1)\sigma^2}{\mu\varepsilon G}$	$\kappa\zeta^2$

[(1)] The rate is derived under the strong convexity assumption. Strong convexity implies the PŁ-condition, but the converse is not true: there exist non-convex PŁ functions Karimi et al. (2016).
[(2)] For comparison, the complexity results of Byz-VR-MARINA and Byz-EF21 are derived by exploring the relationship between (δ,c)-agnostic robust aggregator and (B,κ)-robust aggregator. See Remark 3.5 for details.

2 PRELIMINARIES

We consider a distributed learning system comprising a central server and n workers, denoted as the set $[n] = \mathcal{G} \cup \mathcal{B}$. In this setup, $\mathcal{G}$ represents the subset of reliable or honest workers and $G = |\mathcal{G}|$, while $\mathcal{B}$ consists of malicious or Byzantine workers and $B = |\mathcal{B}|$. Notably, the identities of the honest workers and Byzantine workers are unknown beforehand. The Byzantine workers are assumed to be omniscient Baruch et al. (2019) and capable of colluding with each other to send arbitrary malicious messages to the server. The primary objective is to find the optimal solution to this *distributed stochastic optimization problem*

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

where $f_i(x) = \mathbb{E}_{\xi_i\sim\mathcal{D}_i} f_i(x,\xi_i), \forall i \in \mathcal{G}$. $x \in \mathbb{R}^d$ represents the model parameters to be optimized, while $f_i(x)$ denotes the (typically nonconvex) loss function of the model parameterized by x on the dataset $\mathcal{D}_i$ held by client i. We allow the distributions of malicious nodes, $\mathcal{D}_1,\ldots,\mathcal{D}_n$, to vary arbitrarily. Our objective is to solve the optimization problem (1) in the presence of arbitrary malicious messages sent by Byzantine workers, while ensuring communication efficiency.

The following assumptions will be used throughout the analysis of our algorithms.

Assumption 2.1 (L-Smoothness)**.** *We assume that function $f:\mathbb{R}^d\to\mathbb{R}$ is L-smooth, meaning that for all $x,y\in\mathbb{R}^d$, the following inequality holds:*

$$\|\nabla f_i(x)-\nabla f_i(y)\| \leq L_i\|x-y\|, \tag{2}$$

and

$$\|\nabla f(x)-\nabla f(y)\| \leq L\|x-y\|. \tag{3}$$

Assumption 2.2 (Individual Smoothness)**.** *For each $i=1,\ldots,n$, every realization of $\xi_i\sim\mathcal{D}_i$, the stochastic gradient $\nabla f_i(x,\xi_i)$ is ℓ_i-Lipschitz, i.e., for all $x,y\in\mathbb{R}^d$, we have:*

$$\|\nabla f_i(x,\xi_i)-\nabla f_i(y,\xi_i)\| \leq \ell_i\|x-y\|. \tag{4}$$

We denote the averaged smoothness constants as $\widetilde{L}^2 = G^{-1}\sum_{i\in\mathcal{G}}L_i^2$ and $\widetilde{\ell}^2 = G^{-1}\sum_{i\in\mathcal{G}}\ell_i^2$. Finally, we assume that f is lower bounded, i.e., $f^* := \min_{x\in\mathbb{R}^d} f(x) \geq -\infty$.

In scenarios with arbitrary heterogeneity, distinguishing between regular and Byzantine workers becomes infeasible. Therefore, we adopt a common assumption regarding the heterogeneity of the gradient of local loss functions.

Assumption 2.3 (ζ^2-heterogeneity)**.** *We assume that good workers have ζ^2-heterogeneous local loss functions for some $\zeta\geq 0$, i.e.,*

$$\frac{1}{G}\sum_{i\in\mathcal{G}}\|\nabla f_i(x)-\nabla f(x)\|^2 \leq \zeta^2, \forall x\in\mathbb{R}^d. \tag{5}$$

To model the stochastic noise introduced at each honest worker, we adopt the following assumption.

Assumption 2.4 (Bounded variance (BV)). *There exists $\sigma>0$ such that*

$$\mathbb{E}[\|\nabla f_i(x,\xi_i)-\nabla f_i(x)\|^2]\leq\sigma^2,\forall x\in\mathbb{R}^d\,, \quad (6)$$

where $\mathbb{E}[\cdot]$ is defined over the randomness of the algorithm and $\xi_i\sim\mathcal{D}_i$ are i.i.d. random samples for each $i\in\mathcal{G}$.

We define a Byzantine-robust algorithm as an algorithm guaranteed to find an ε-approximate stationary point for f_i despite the presence of B Byzantine workers. In particular, we introduce the formal definition of Byzantine robustness as follows.

Definition 2.5 ((B,ε)-Byzantine robustness). *A learning algorithm is said (B,ε)-Byzantine robust if, even in the presence of B Byzantine workers, it outputs $\hat{x}$ satisfying*

$$\mathbb{E}[\|\nabla f(\hat{x})\|^2]\leq\varepsilon\,. \quad (7)$$

Achieving (B,ε)-Byzantine robustness is generally not feasible for any ε when the number of Byzantine workers B is at least half the total workers ($B\geq n/2$). Therefore, in this work, we assume an upper bound on B, specifically $B<n/2$, to ensure robustness.

Several robust aggregation methods have been proposed, including the coordinate-wise trimmed mean (CWTM) Yin et al. (2018) and centered clipping Karimireddy et al. (2021). To formally assess the robustness of aggregation techniques, we adopt the concept of (B,κ)-robustness Allouah et al. (2023). This property ensures that for any subset of inputs of size G, the output of the aggregation rule remains close to the average of those inputs. This concept serves as a useful metric for comparing the robustness of different aggregation methods. For a detailed analysis and formal quantification of commonly used aggregation rules, refer to Allouah et al. (2023).

Definition 2.6 ((B,κ)-robustness). *Given an integer $B<n/2$ and a real number $\kappa\geq 0$, an aggregation rule F is (B,κ)-robust if for any set of n vectors $\{g_1,g_2,\ldots,g_n\}$, and any subset $S\subseteq[n]$ with $|S|=G$,*

$$\|F(g_1,g_2,\ldots,g_n)-\overline{g}_S\|^2\leq\frac{\kappa}{|S|}\sum_{i\in S}\|g_i-\overline{g}_S\|^2\,, \quad (8)$$

where $\overline{g}_S:=G^{-1}\sum_{i\in S}g_i$.

To facilitate communication-efficient learning, we introduce compression techniques for message transmission. A general compression operator is defined as follows.

Definition 2.7 (Contractive compressors). *A (possibly randomized) mapping $\mathcal{C}:\mathbb{R}^d\to\mathbb{R}^d$ is called a contractive compression operator if there exists a constant $\alpha\in(0,1]$ such that*

$$\mathbb{E}[\|\mathcal{C}(x)-x\|^2]\leq(1-\alpha)\|x\|^2,\ \forall x\in\mathbb{R}^d. \quad (9)$$

In this paper, we focus on the Top_k sparsification-based compression operator. This operator sorts the input vector, selects the largest k elements, and transmits these elements along with their original indices. Another commonly used sparsification operator is Rand_k, which randomly sets $d-k$ elements of the input vector to zero, where k is an integer between 1 and d. This approach achieves a sparsity ratio of d/k. For a detailed overview of both biased and unbiased compressors, please refer to Beznosikov et al. (2023).

3 BYZANTINE-ROBUST DISTRIBUTED LEARNING

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

3.1 Byzantine-DM21

We summarize our new method Byz-DM21 in Algorithm 1. At each iteration of Byz-DM21, the model parameter $x^{(t+1)}$ is updated by the parameter server using the formula: $x^{(t+1)}=x^{(t)}-\gamma g^{(t)}$, where γ represents the step size and $g^{(t)}$ is the gradient estimator received from the parameter server. Following the update, the server broadcasts the updated model $x^{(t+1)}$ to all workers. Upon receiving $x^{(t+1)}$, the good workers proceed with the following steps: Update their first momentum estimate $v_i^{(t+1)}$ and the second momentum estimate $u_i^{(t+1)}$(lines 5 and 6). The change in $g_i^{(t)}$ is then compressed using the expression: $c_i^{(t+1)}=\mathcal{C}(u_i^{(t+1)}-g_i^{(t)})$, these compressed vectors are sent back to the server (line 7). Simultaneously, the honest workers update their local state according to: $g_i^{(t+1)}=g_i^{(t)}+c_i^{(t+1)}$ reflecting the adjustments made due to compression (line 8). After that, the server gathers the results of computations from the workers and applies a (B,κ)-robust aggregator to compute the next estimator $g^{(t+1)}$.

Two crucial aspects of the Byz-DM21 are introduced as follows.

First, honest workers transmit only compressed updates of their local momentum variables and estimated momentum variables to the server. This approach not only reduces communication overhead but also helps to identify and exclude Byzantine workers who attempt to subvert the algorithm by transmitting dense, inconsistent vectors. *Second*, the double momentum method enhances the efficiency of the optimization process by combining first and second momentum estimations. On the one hand, the first momentum helps smooth gradient fluctuations, reduces the influence of noise, and accelerates convergence. On the other hand, the second momentum further optimizes the update direction, prevents gradient oscillations, and improves local con-

Algorithm 1 Byzantine-DM21

Input: initial model $x^{(0)}$, step-size $\gamma > 0$, momentum coefficient $\eta \in (0, 1]$, robust aggregation F, initial batch size b, the number of rounds T

Initialization: for every honest worker $i \in \mathcal{G}$, $v_i^{(0)} = u_i^{(0)} = g_i^{(0)} = \nabla f_i(x^{(0)}, \xi_i^{(0)})$, each worker $i \in [n]$ sends $g_i^{(0)}$ to the server, $g^{(0)} = F(\{g_1^{(0)}, \ldots, g_n^{(0)}\})$

for $t = 0, 1, \ldots, T-1$ **do**
  Server computes $x^{(t+1)} = x^{(t)} - \gamma g^{(t)}$ and broadcasts $x^{(t+1)}$ to all workers
  for every honest worker $i \in \mathcal{G}$ in parallel **do**
    Compute the first momentum estimator:
    $$v_i^{(t+1)} = \begin{cases} (1-\eta)v_i^{(t)} + \eta \nabla f_i(x^{(t+1)}, \xi_i^{(t+1)}) & \text{(Byz-DM21)} \\ \nabla f_i(x^{(t+1)}, \xi_i^{(t+1)}) + (1-\eta)(v_i^{(t)} - \nabla f_i(x^{(t)}, \xi_i^{(t+1)})) & \text{(Byz-VR-DM21)} \end{cases}$$
    Compute the second momentum estimator: $u_i^{(t+1)} = (1-\eta)u_i^{(t)} + \eta(v_i^{(t+1)})$
    Compress $c_i^{(t+1)} = \mathcal{C}(u_i^{(t+1)} - g_i^{(t)})$ and send $c_i^{(t+1)}$ to the server
    Update local state $g_i^{(t+1)} = g_i^{(t)} + c_i^{(t+1)}$
  end for
  Server updates $g^{(t+1)} = F(\{g_1^{(t+1)}, \ldots, g_n^{(t+1)}\})$ via $g_i^{(t+1)} = g_i^{(t)} + c_i^{(t+1)}, \forall i \in [n]$
end for
Return: A random $\hat{x}^{(T)}$ from $\{x^{(t)}\}_{t=0}^{T-1}$

vergence. This double momentum mechanism not only accelerates the training process but also reduces the impact of gradient noise and enhances robustness.

3.2 Convergence of Byz-DM21 for General Non-Convex Problems

The convergence analysis of Byz-DM21 hinges on the monotonicity of the following Lyapunov function:

$$\Phi^{(t)} = \delta_t + \frac{4\gamma}{\eta}\|\widetilde{M}^{(t)}\|^2 + \Big(\frac{48\gamma(8\kappa+1)(\eta^4+6\eta^2)}{\eta\alpha^2 G} + \frac{8\gamma(28\kappa+3)}{\eta G}\Big)\sum_{i\in\mathcal{G}}\|M_i^{(t)}\|^2, \tag{10}$$

where $\delta_t = f(x^{(t)}) - f(x^*), M_i^{(t)} = v_i^{(t)} - \nabla f_i(x^{(t)})$, and $\widetilde{M}^{(t)} = G^{-1}\sum_{i\in\mathcal{G}} v_i^{(t)} - \nabla f_{\mathcal{G}}(x^{(t)})$. The primary convergence result for general non-convex functions is articulated in Theorem 3.1, with the corresponding proof detailed in Appendix F.

Theorem 3.1. *Assuming that Assumptions 2.1, 2.3, and 2.4 hold, we consider Algorithm 1 for solving the distributed learning problem* (1) *with $B<n/2$ Byzantine workers and communication compression characterized by the parameter $\alpha \in (0, 1]$ as per Definition 2.7. If $\eta \le 1$, and $\gamma \le \min\Big\{\frac{\alpha}{4\widetilde{L}\sqrt{234(8\kappa+1)}}, \frac{\eta}{4\sqrt{(56\kappa+6)\widetilde{L}^2+L^2}}\Big\}$, then*

$$\mathbb{E}\Big[\|\nabla f(\hat{x}^{(T)})\|^2\Big] \le \frac{\Phi^{(0)}}{\gamma T} + \Big(\frac{48(\eta^5+7\eta^3)(8\kappa+1)}{\alpha^2} + 4\eta(64\kappa+7) + \frac{4\eta}{G} + \frac{8(8\kappa+1)\eta^4}{\alpha}\Big)\sigma^2 + 32\kappa\zeta^2,$$

where $\hat{x}^{(T)}$ is sampled uniformly at random from the iterations of the method, $\Phi^{(0)}$ is defined in (10)*. Then choice $\eta \le \mathcal{O}\Big(\min\Big\{\Big(\frac{\alpha^2\delta_0\widehat{L}}{(\kappa+1)\sigma^2 T}\Big)^{1/6}, \Big(\frac{\alpha^2\delta_0\widehat{L}}{(\kappa+1)\sigma^2 T}\Big)^{1/4}, \Big(\frac{\delta_0\widehat{L}}{(\kappa+1)\sigma^2 T}\Big)^{1/2}, \Big(\frac{\delta_0 G\widehat{L}}{\sigma^2 T}\Big)^{1/2}, \Big(\frac{\alpha\widehat{L}\delta_0}{(\kappa+1)\sigma^2 T}\Big)^{1/5}\Big\}\Big)$, where $\widehat{L} \overset{def}{=} 8\sqrt{(56\kappa+6)\widetilde{L}^2 + L^2}$, we obtain*

$$\mathbb{E}\Big[\|\nabla f(\hat{x}^{(T)})\|^2\Big] = \mathcal{O}\Big(\Big(\frac{(\kappa+1)^{1/5}\sigma^{2/5}\widehat{L}\delta_0}{\alpha^{2/5}T}\Big)^{5/6} + \kappa\zeta^2 + \Big(\frac{(\kappa+1)^{1/3}\sigma^{2/3}\widehat{L}\delta_0}{\alpha^{2/3}T}\Big)^{3/4} + \Big(\frac{(\kappa+1)\sigma^2\widehat{L}\delta_0}{T}\Big)^{1/2} + \Big(\frac{(\kappa+1)^{1/4}\sigma^{1/2}\widehat{L}\delta_0}{\alpha^{1/4}T}\Big)^{4/5} + \frac{\Phi^{(0)}}{\gamma T} + \Big(\frac{\sigma^2\widehat{L}\delta_0}{GT}\Big)^{1/2}\Big).$$

Theorem 3.1 establishes that the proposed algorithm, Byz-DM21, converges in $\mathbb{E}[\|\nabla f(\hat{x}^T)\|^2]$, adhering to the Byzantine robustness criterion specified in Definition 2.5. Consequently, Byz-DM21 achieves a convergence rate comparable to that of standard SGD with momentum Cutkosky and Mehta (2020). Furthermore, prior studies Cutkosky and Mehta (2020); Arjevani et al. (2023) have shown that, under standard Assumptions, the convergence rate of $\mathcal{O}(1/T^{1/2})$ is optimal for SGD. Although Byz-VR-MARINA Gorbunov et al. (2023) attains a faster convergence rate by intermittently using full gradients, this approach incurs significant computational costs, particularly in real-world scenarios involving large-scale training data.

Remark 3.2. *Omitting constants and higher-order terms, Theorem 3.1 gives $\mathbb{E}[\|\nabla f(\hat{x}^{(T)})\|^2] \lesssim \frac{\sigma}{\sqrt{GT}} + \frac{\sqrt{\kappa}\sigma}{\sqrt{T}}$. When there is no Byzantine adversary, i.e., $B = 0$, employing the standard mean aggregator (which*

is (0,0)-robust) yields a convergence rate of $\mathcal{O}\left(\frac{\sigma}{\sqrt{GT}}\right)$, *which improves with the number of workers* G. *Additionally,* Byz-DM21 *attains better sample complexity than* Byz-EF21-SGDM *because its bound in Theorem* (3.1) *involves* η^4/α, *which is more favorable compared to the terms* η^2/α *in* Byz-EF21-SGDM. *Consequently, this term becomes dominated by other terms and vanishes in Corollary 3.3.*

Corollary 3.3. *To guarantee* $\mathbb{E}[\|\nabla f(\hat{x}^{(T)})\|^2] \leq \varepsilon^2$ *for* $\varepsilon^2 \geq 64\kappa\zeta^2$, *we obtain*

$$T = \mathcal{O}\Big(\frac{\widetilde{L}\sqrt{\kappa+1}}{\alpha\varepsilon^2} + \frac{(\kappa+1)^{1/5}\sigma^{2/5}\widehat{L}}{\alpha^{2/5}\varepsilon^{12/5}} + \frac{(\kappa+1)^{1/4}\sigma^{1/2}\widehat{L}}{\alpha^{1/4}\varepsilon^{5/2}} + \frac{(\kappa+1)^{1/3}\sigma^{2/3}\widehat{L}}{\alpha^{2/3}\varepsilon^{8/3}} + \frac{(G(\kappa+1)+1)\sigma^2\widehat{L}}{G\varepsilon^4}\Big).$$

Next, we consider a special case where local full gradients are available to workers, i.e., $\sigma = 0$.

Corollary 3.4. *If* $\sigma = 0$, *then* $\mathbb{E}\left[\|\nabla f(\hat{x}^{(T)})\|\right] \leq \varepsilon$ *after* $T = \mathcal{O}(\tilde{L}\sqrt{\kappa+1}/\alpha\varepsilon^2)$ *iterations.*

Remark 3.5. *In the special case of full gradients, Rammal et al. (2024) established a complexity of* $\mathcal{O}(1+\sqrt{c\delta}/\alpha\varepsilon^2)$, *where* $\delta = B/n$ *and* c *are parameters characterizing the agnostic robust aggregator (ARAgg) Karimireddy et al. (2021). Note that an* (B,κ)*-robust aggregation rule also qualifies as a* (δ, c)*-ARAgg with* $c = \kappa n/2B$ *Allouah et al. (2023). As a result, Corollary 3.4 provides a slight improvement over the complexity bound* $\mathcal{O}(1+\sqrt{\kappa}/\alpha\varepsilon^2)$ *established in Rammal et al. (2024). Moreover, in the absence of Byzantine faults, we adopt the standard mean aggregator, which is (0,0)-robust. Then, the iteration complexity is* $T = \mathcal{O}(\widetilde{L}/(\alpha\varepsilon^2))$, *achieving the lower bound for Byzantine-free distributed learning with communication compression Huang et al. (2022). We note that this asymptotic complexity bound also holds for any constant* κ, *achieved by many other aggregators unless* n *is big and* $B/n \to 1/2$.

3.3 Convergence of Byz-DM21 under the Polyak-Łojasiewicz Condition

In this section, we provide complexity bounds for Byz-DM21 under the Polyak-Łojasiewicz (PŁ) condition.

Assumption 3.6 (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*

$$2\mu(f(x) - f(x^*)) \leq \|\nabla f(x)\|^2, \quad \forall x \in \mathbb{R}, \tag{11}$$

where $f(x^*) = \inf_{x\in\mathbb{R}^d} f(x) > -\infty$. *Here we use a different notion of an* ε*-solution: it is a (random) point* $\hat{x}$, *such that* $\mathbb{E}[f(\hat{x}^{(T)}) - f(x^*)] \leq \varepsilon$. *Under this and previously introduced assumptions, we derive the following result.*

Theorem 3.7. *Let Assumptions 2.1, 2.3, 2.4 and 3.6 be satisfied, then choose momentum* $\eta \leq \mathcal{O}\Big(\min\Big\{\frac{\mu\varepsilon G}{(G(\kappa+1)+1)\sigma^2}, \frac{\mu\alpha^2\varepsilon^{1/3}}{(\kappa+1)\sigma^2}, \frac{\mu\alpha\varepsilon^{1/4}}{(\kappa+1)\sigma^2}\Big\}\Big)$, *after*

$$T = \mathcal{O}\Big(\frac{(G(\kappa+1)+1)\sigma^2}{\mu\varepsilon G} + \frac{(\kappa+1)\sigma^2}{\mu\alpha^2\varepsilon^{1/3}} + \frac{(\kappa+1)\sigma^2}{\mu\alpha\varepsilon^{1/4}} + \frac{L}{\mu} + \frac{L\sigma^2(G(\kappa+1)+1)\sqrt{(\kappa+1)}}{\mu^2\varepsilon G}\Big)$$

iterations with stepsize $\gamma \leq \min\Big\{\frac{\eta}{2\mu}, \frac{\alpha}{4\mu}, L^{-1}\Big(1 + \sqrt{\frac{104\alpha^2(8\kappa+1)+48\eta^2(8\kappa+1)(2\eta^4+28\eta^2+\alpha^2+48)}{\alpha^2\eta^2}}\Big)^{-1}\Big\}$, Byz-DM21 *produces a point* $\hat{x}^{(T)}$ *for which* $\mathbb{E}[f(\hat{x}^{(T)}) - f(x^*)] \leq \varepsilon$.

4 INCORPORATING VARIANCE REDUCTION

In this work, we employ a gradient estimator inspired by those utilized in Cutkosky and Orabona (2019) and Tran-Dinh et al. (2019). The proposed gradient estimator of each node integrates the advantages of the widely-used SARAH estimator Nguyen et al. (2017) and the unbiased SGD gradient estimator. Formally, the gradient estimator of node $i \in \mathcal{G}$ at time step t is expressed as:

$$v_i^{(t)} = \eta \underbrace{\nabla f_i(x^{(t)}, \xi_i^{(t)})}_{\text{SGD}} + (1-\eta)\underbrace{\Big(v_i^{(t-1)} + \nabla f_i(x^{(t)}, \xi_i^{(t)}) - \nabla f_i(x^{(t-1)}, \xi_i^{(t)})\Big)}_{\text{SARAH}}, \tag{12}$$

with the parameter $\eta \in (0, 1]$ as the momentum parameter. Next, we discuss the algorithm.

We extend Byz-DM21 to Byz-VR-DM21, incorporating local variance reduction on all nodes to progressively eliminate variance from stochastic approximations. This new algorithm is depicted in Algorithm 1. Its main difference from the original Byz-DM21 is that in the momentum update rule (Line 5), an additional term of $(1-\eta)(\nabla f(x^{(t+1)}, \xi_i^{(t+1)}) - \nabla f(x^{(t)}, \xi_i^{(t+1)}))$ is added, inspired by the STORM algorithm Cutkosky and Orabona (2019). This term corrects the bias of $v_i^{(t)}$ so that it is an unbiased estimate of $\nabla f(x^{(t)})$ in the current iteration condition, i.e., $\mathbb{E}[v_i^{(t)}|x^{(t)}] = \nabla f_i(x^{(t)})$. We will also show that it reduces the variance and accelerates the convergence.

4.1 Convergence of Byz-VR-DM21 for General Non-Convex Problems

We now present a variant of Byz-VR-DM21, outlined in Algorithm 1. The convergence analysis of this method

hinges on the monotonicity of the following Lyapunov function:

$$\Phi^{(t)} = \delta_t + \frac{4\gamma}{\eta}\|\widetilde{M}^{(t)}\|^2 + \left(\frac{48\gamma(8\kappa+1)(\eta^3+6\eta)}{\alpha^2 G} + \frac{8\gamma(28\kappa+3)}{\eta G}\right)\sum_{i\in\mathcal{G}}\|M_i^{(t)}\|^2, \tag{13}$$

where $\delta_t = f(x^{(t)}) - f(x^*), M_i^{(t)} = v_i^{(t)} - \nabla f_i(x^{(t)})$, and $\widetilde{M}^{(t)} = G^{-1}\sum_{i\in\mathcal{G}} v_i^{(t)} - \nabla f_{\mathcal{G}}(x^{(t)})$. The primary convergence result for general non-convex functions is articulated in Theorem 4.1, with the corresponding proof detailed in Appendix G.

Theorem 4.1. *Let Assumptions 2.1, 2.2, 2.3 and 2.4 be satisfied, consider Algorithm 1 for solving the distributed learning problem (1) with $B < n/2$ Byzantine workers and communication compression characterized by the parameter $\alpha \in (0,1]$ as per Definition 2.7. If $\eta \le 1$, and $\gamma \le \min\left\{\frac{\alpha}{10\sqrt{(8\kappa+1)(49\widetilde{\ell}^2+15\widetilde{L}^2)}}, \frac{\sqrt{\eta}}{20\sqrt{(8\kappa+1)\widetilde{\ell}^2}}\right\}$, then*

$$\mathbb{E}\left[\|\nabla f(\hat{x}^{(T)})\|^2\right] \le \frac{\Phi^{(0)}}{\gamma T} + \left(\frac{96(8\kappa+1)(\eta^5+7\eta^3)}{\alpha^2} + 8\eta(60\kappa+7) + \frac{8\eta}{G} + \frac{16\eta^4(8\kappa+1)}{\alpha}\right)\sigma^2 + 32\kappa\zeta^2,$$

where $\hat{x}^{(T)}$ is sampled uniformly at random from the iterates of the method. By setting $\eta \le \mathcal{O}\Big(\min\Big\{\Big(\frac{\alpha^2\delta_0\widehat{\ell}}{(\kappa+1)\sigma^2 T}\Big)^{2/11}, \Big(\frac{\alpha^2\delta_0\widehat{\ell}}{(\kappa+1)\sigma^2 T}\Big)^{2/7}, \Big(\frac{\delta_0\widehat{\ell}}{(\kappa+1)\sigma^2 T}\Big)^{2/3}, \Big(\frac{\delta_0 G\widehat{\ell}}{\sigma^2 T}\Big)^{2/3}, \Big(\frac{\alpha\delta_0\widehat{\ell}}{(\kappa+1)\sigma^2 T}\Big)^{2/9}\Big\}\Big)$, where $\widehat{\ell} \overset{def}{=} 40\sqrt{(8\kappa+1)\widetilde{\ell}^2}$, we obtain

$$\mathbb{E}\left[\|\nabla f(\hat{x}^{(T)})\|^2\right] = \mathcal{O}\Big(\Big(\frac{(\kappa+1)^{1/10}\sigma^{2/10}\delta_0\widehat{\ell}}{\alpha^{2/10}T}\Big)^{10/11} + \frac{\Phi^{(0)}}{\gamma T} + \kappa\zeta^2 + \Big(\frac{(\kappa+1)^{1/6}\sigma^{2/6}\delta_0\widehat{\ell}}{\alpha^{2/6}T}\Big)^{6/7} + \Big(\frac{\sigma\delta_0\widehat{\ell}}{G^{1/2}T}\Big)^{2/3} + \Big(\frac{(\kappa+1)^{1/2}\sigma\delta_0\widehat{\ell}}{T}\Big)^{2/3} + \Big(\frac{(\kappa+1)^{1/8}\sigma^{1/4}\delta_0\widehat{\ell}}{\alpha^{1/8}T}\Big)^{8/9}\Big).$$

Corollary 4.2. *To guarantee $\mathbb{E}[\|\nabla f(\hat{x}^{(T)})\|^2] \le \varepsilon^2$ for $\varepsilon^2 \ge 64\kappa\zeta^2$, then*

$$T = \mathcal{O}\Big(\frac{(G(\kappa+1)^{1/2}+1)\sigma\widehat{\ell}}{G^{1/2}\varepsilon^3} + \frac{(\kappa+1)^{1/10}\sigma^{1/5}\widehat{\ell}}{\alpha^{2/5}\varepsilon^{11/5}} + \frac{(\kappa+1)^{1/6}\sigma^{1/3}\widehat{\ell}}{\alpha^{1/3}\varepsilon^{7/3}} + \frac{(\kappa+1)^{1/8}\sigma^{1/4}\widehat{\ell}}{\alpha^{1/8}\varepsilon^{9/4}} + \frac{\widetilde{L}\sqrt{\kappa+1}}{\alpha\varepsilon^2}\Big).$$

Corollary 4.3. *If $\sigma = 0$, then $\mathbb{E}\left[\|\nabla f(\hat{x}^{(T)})\|\right] \le \varepsilon$ after $T = \mathcal{O}(\widetilde{L}\sqrt{\kappa+1}/\alpha\varepsilon^2)$ iterations.*

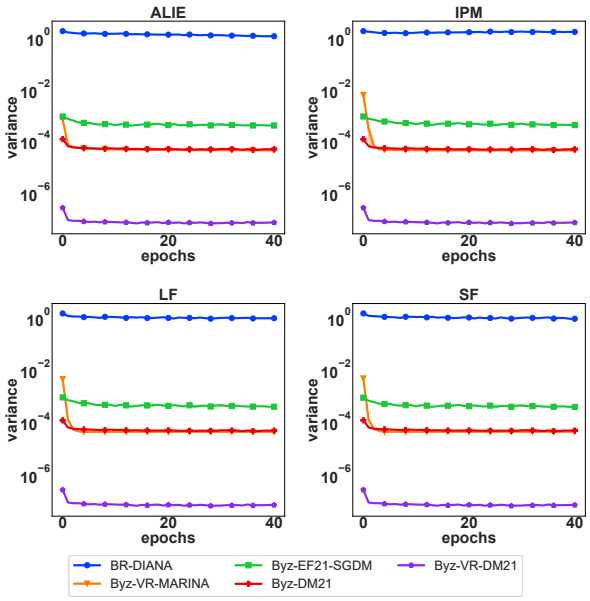

Figure 1: The training variance of honest messages under four attack scenarios on the a9a dataset.

Remark 4.4. *We emphasize several key properties of Theorem 3.1 and Theorem 4.1. These theorems provide the theoretical guarantee for the convergence of the error feedback method with stochastic gradients in the presence of Byzantine attacks. In the heterogeneous case (i.e., $\zeta > 0$), the algorithm does not ensure that $\mathbb{E}\left[\|\nabla f(\hat{x}^{(T)})\|\right]$ can be made arbitrarily small. This limitation is characteristic of all Byzantine-robust algorithms in heterogeneous environments. Specifically, with an order-optimal robustness coefficient $\kappa = \mathcal{O}(B/n)$, as achieved by* CWTM*, the result is consistent with the lower bound $\Omega(B/n\zeta^2)$ established by Allouah et al. (2023). Moreover, the highest attainable accuracy of* Byz-DM21 *and* Byz-VR-DM21 *is tighter than that of* Byz-VR-MARINA *and* Byz-EF21 *(see Table 1).*

4.2 Convergence of Byz-VR-DM21 under the Polyak-Łojasiewicz Condition

Theorem 4.5. *Let Assumptions 2.1, 2.2, 2.3, 2.4 and 3.6 be satisfied, then choose momentum $\eta \le \mathcal{O}\Big(\min\Big\{\frac{\mu G\varepsilon}{(G(\kappa+1)+1)\sigma^2}, \frac{\mu\alpha^2\varepsilon^{1/3}}{(\kappa+1)\sigma^2}, \frac{\mu\alpha\varepsilon^{1/4}}{(\kappa+1)\sigma^2}\Big\}\Big)$, after*

$$T = \mathcal{O}\Big(\frac{(G(\kappa+1)+1)\sigma^2}{\mu\varepsilon G} + \frac{(\kappa+1)\sigma^2}{\mu\alpha^2\varepsilon^{1/3}} + \frac{(\kappa+1)\sigma^2}{\mu\alpha\varepsilon^{1/4}} + \frac{L}{\mu} + \frac{L\sigma^2(G(\kappa+1)+1)}{\mu^2\varepsilon G}\Big)$$

iterations with stepsize $\gamma \le \min\Big\{\Big(L + 4\sqrt{8\kappa+1}\sqrt{\frac{16\alpha^2(L^2+7\eta\ell^2)+\eta^2(L^2(3\eta^2+24)+\ell^2(12\eta^3+\alpha\eta^2+156\eta))}{\eta^2\alpha^2}}\Big)^{-1},$

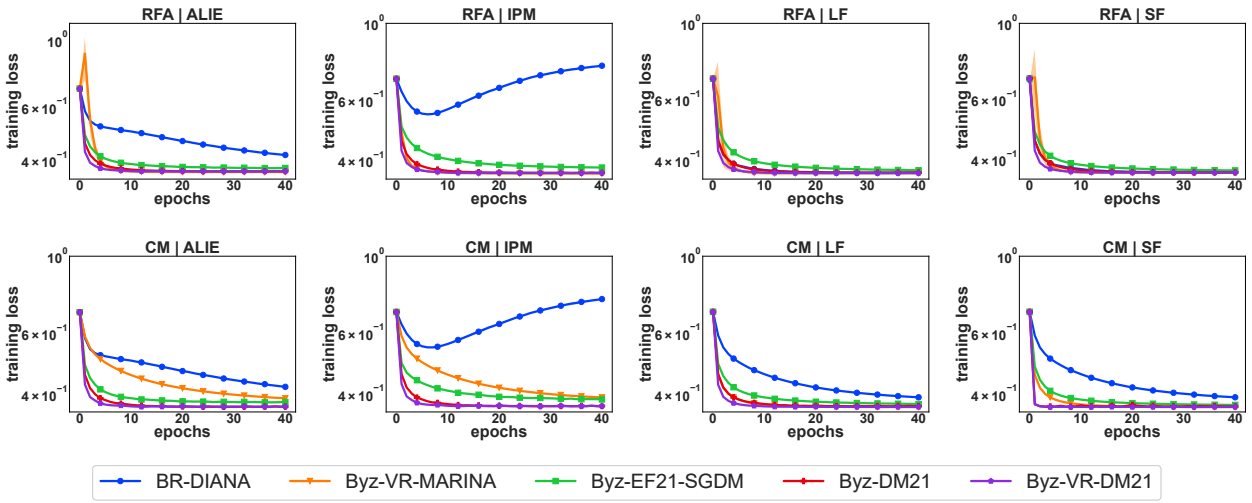

Figure 2: The training loss of RFA and CM under four attack scenarios (SF, IPM, LF, ALIE) on the a9a dataset in a heterogeneous setting. We use $k = 0.1d$ for both Rand_k and Top_k compressors.

$\frac{\eta}{2\mu}, \frac{\alpha}{4\mu}\Big\}$, Byz-VR-DM21 *produces a point* $\hat{x}^{(T)}$ *for which* $\mathbb{E}[f(\hat{x}^{(T)}) - f(x^*)] \leq \varepsilon$.

5 NUMERICAL EXPERIMENTS

In this section, we demonstrate the performance of the proposed method. The goal of our experimental evaluation is to showcase the benefits of double momentum to mitigate Byzantine workers. We consider two binary classification tasks: a9a and w8a from the LIBSVM Chang and Lin (2011) dataset, and image classification on CIFAR-10 Krizhevsky et al. (2009) and FEMNIST Caldas et al. (2018). Due to space limitations, we present results only on a9a and defer the rest to the Appendix D.

Adversarial attacks. We address a binary logistic regression problem with a regularizer, using the a9a dataset from LIBSVM Chang and Lin (2011). The data is distributed across $n = 20$ workers, 8 of which are Byzantine. For aggregation, we use Robust Federated Averaging (RFA) Pillutla et al. (2022), Coordinate-wise Median (CM) Yin et al. (2018), and the NNM algorithm Allouah et al. (2023). The experiments test four Byzantine attack strategies: Sign Flipping (SF) Allen-Zhu et al. (2021), Label Flipping (LF) Allen-Zhu et al. (2021), Inner Product Manipulation (IPM) Xie et al. (2020), and A Little Is Enough (ALIE) Baruch et al. (2019) (details in Appendix C). We compare our algorithm with BR-DIANA [2] Mishchenko et al. (2019), Byz-VR-MARINA Gorbunov et al. (2023), and Byz-EF21-SGDM Liu et al. (2026). For the contractive compressor, we use Top_k, while Byz-EF21-SGDM uses it too, and all other algorithms use Rand_k [3].

Empirical results. Figure 1 illustrates the training variance of honest messages for the compared algorithms. The results demonstrate that Byz-VR-DM21 effectively reduces the variance of the stochastic gradient, maintaining a consistently low level of variance even after convergence. This highlights Byz-VR-DM21's enhanced robustness in mitigating noise and interference from Byzantine nodes. On the other hand, Byz-DM21 achieves a comparable variance level to Byz-VR-MARINA despite not employing explicit variance reduction techniques, showcasing the advantages conferred by its double-momentum design. Figure 2 depicts the training loss across the four methods under various attack scenarios. Both Byz-DM21 and Byz-VR-DM21 exhibit rapid convergence and strong resilience to Byzantine interference, outperforming the other algorithms. In contrast, Byz-VR-MARINA, although capable of quick convergence, suffers from significant fluctuations when exposed to Byzantine attacks. Notably, Byz-VR-DM21 achieves a faster convergence rate compared to Byz-DM21, which can be attributed to the effectiveness of its variance reduction mechanism.

Reproducibility. To ensure reproducibility, all experiments were conducted using three different random seeds. We report the mean training loss along with one standard error.

[2] BR-DIANA is a version of BROADCAST with the SGD estimator instead of the SAGA estimator. BROADCAST consumes a large amount of memory, which scales linearly with the number of data points. We compare Byrd-SAGA in Appendix D.

[3] To ensure a fair comparison, each method employs a theoretically compatible compressor.

6 CONCLUSION

We introduce Byz-DM21 and Byz-VR-DM21, Byzantine-tolerant schemes designed to harness the empirical benefits of double momentum and variance reduction, while preserving communication efficiency. Unlike most existing Byzantine-tolerant methods, our new algorithm leverages stochastic gradients and is batch-free, meaning it does not require computing full gradients or additional tuning. Through theoretical analysis, we prove that our new algorithm has tight lower bounds and a smaller neighborhood size. It matches the upper bound results in both stochastic and full gradient scenarios when the problem is Byzantine-free, and also converges to a smaller neighborhood around the optimal solution. We further show that, under the PŁ condition, our algorithm admits a faster convergence guarantee in terms of the optimality gap. Moreover, we demonstrate the robustness of our algorithm against adversarial attacks through experiments on binary and image classification tasks.

Acknowledgements

Yanghao Li and Yuhao Yi acknowledge support from the National Natural Science Foundation of China under Grant 62303338, while Changxin Liu acknowledges support under Grant 62573196.

References

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. (2018). The convergence of sparsified gradient methods. *Advances in Neural Information Processing Systems*, 31.

Allen-Zhu, Z. (2018). Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(221):1–51.

Allen-Zhu, Z., Ebrahimianghazani, 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):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.

Bernstein, J., Zhao, J., Azizzadenesheli, K., and Anandkumar, A. (2019). signSGD with majority vote is communication efficient and fault tolerant. In *International Conference on Learning Representations*.

Beznosikov, A., Horváth, S., Richtárik, P., and Safaryan, M. (2023). On biased compression for distributed learning. *Journal of Machine Learning Research*, 24(276):1–50.

Blanchard, P., El Mhamdi, E. M., Guerraoui, R., and Stainer, J. (2017). Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 30.

Caldas, S., Duddu, S. M. K., Wu, P., Li, T., Konečnỳ, J., McMahan, H. B., Smith, V., and Talwalkar, A. (2018). Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*.

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, T., Giannakis, G., Sun, T., and Yin, W. (2018). Lag: Lazily aggregated gradient for communication-efficient distributed learning. *Advances in neural information processing systems*, 31.

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.

Cohen, G., Afshar, S., Tapson, J., and Van Schaik, A. (2017). Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE.

Cutkosky, A. and Mehta, H. (2020). Momentum improves normalized sgd. In *International conference on machine learning*, pages 2260–2268. PMLR.

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

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.

Diakonikolas, I. and Kane, D. M. (2019). Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*.

El Mhamdi, E. M., Guerraoui, R., and Rouault, S. L. A. (2021). Distributed momentum for byzantine-resilient stochastic gradient descent. In *9th International Conference on Learning Representations (ICLR)*.

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.

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*.

Fatkhullin, I., Tyurin, A., and Richtárik, P. (2023). Momentum provably improves error feedback! In *Thirty-seventh Conference on Neural Information Processing Systems*.

Fedin, N. and Gorbunov, E. (2023). Byzantine-robust loopless stochastic variance-reduced gradient. In *International Conference on Mathematical Optimization Theory and Operations Research*, pages 39–53. Springer.

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. In *The Eleventh International Conference on Learning Representations*.

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

Guerraoui, R., Gupta, N., and Pinot, R. (2024). Byzantine machine learning: A primer. *ACM Computing Surveys*, 56(7):1–39.

Haddadpour, F., Kamani, M. M., Mahdavi, M., and Cadambe, V. (2019). Trading redundancy for communication: Speeding up distributed sgd for non-convex optimization. In *International Conference on Machine Learning*, pages 2545–2554. PMLR.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hong, S., Yang, H., and Lee, J. (2022). Hierarchical group testing for byzantine attack identification in distributed matrix multiplication. *IEEE Journal on Selected Areas in Communications*, 40(3):1013–1029.

Horvóth, S., Ho, C.-Y., Horvath, L., Sahu, A. N., Canini, M., and Richtárik, P. (2022). Natural compression for distributed deep learning. In *Mathematical and Scientific Machine Learning*, pages 129–141. PMLR.

Huang, X., Chen, Y., Yin, W., and Yuan, K. (2022). Lower bounds and nearly optimal algorithms in distributed learning with communication compression. *Advances in Neural Information Processing Systems*, 35:18955–18969.

Jaggi, M., Smith, V., Takác, M., Terhorst, J., Krishnan, S., Hofmann, T., and Jordan, M. I. (2014). Communication-efficient distributed dual coordinate ascent. *Advances in neural information processing systems*, 27.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26.

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.

Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pages 795–811. Springer.

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. In *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.

Konstantinidis, K. and Ramamoorthy, A. (2021). Byzshield: An efficient and robust system for distributed training. *Proceedings of Machine Learning and Systems*, 3:812–828.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Lamport, L., Shostak, R., and Pease, M. (2019). The byzantine generals problem. In *Concurrency: the works of leslie lamport*, pages 203–226.

Lan, G., Li, Z., and Zhou, Y. (2019). A unified variance-reduced accelerated gradient method for convex optimization. *Advances in Neural Information Processing Systems*, 32.

Lan, G. and Zhou, Y. (2018). An optimal randomized incremental gradient method. *Mathematical programming*, 171:167–215.

Lee, J. D., Lin, Q., Ma, T., and Yang, T. (2017). Distributed stochastic variance reduced gradient methods by sampling extra data with replacement. *Journal of Machine Learning Research*, 18(122):1–43.

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.

Liu, C., Li, Y., Yi, Y., and Johansson, K. H. (2026). Byzantine-robust and communication-efficient distributed learning via compressed momentum filtering. *IEEE Transactions on Neural Networks and Learning Systems*.

Liu, H., Shan, L., Bao, H., You, R., Yi, Y., and Lv, J. (2025). Lid-fl: Towards list-decodable federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18825–18833.

McMahan, B. and Ramage, D. (2017). Federated learning: Collaborative machine learning without centralized training data. *Google Research Blog*, 3.

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.

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

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

Rammal, A., Gruntkowska, K., Fedin, N., Gorbunov, E., and Richtárik, P. (2024). Communication compression for byzantine robust learning: New efficient algorithms and improved rates. In *International Conference on Artificial Intelligence and Statistics*, pages 1207–1215. PMLR.

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*, 34:4384–4396.

Safaryan, M., Shulgin, E., and Richtárik, P. (2022). Uncertainty principle for communication compression in distributed and federated learning and the search for an optimal compressor. *Information and Inference: A Journal of the IMA*, 11(2):557–580.

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*, 34:8133–8146.

Schmidt, M., Le Roux, N., and Bach, F. (2017). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162:83–112.

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 *Interspeech*, volume 2014, pages 1058–1062. Singapore.

Shi, W., Cao, J., Zhang, Q., Li, Y., and Xu, L. (2016). Edge computing: Vision and challenges. *IEEE internet of things journal*, 3(5):637–646.

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

Tran-Dinh, Q., Pham, N. H., Phan, D. T., and Nguyen, L. M. (2019). Hybrid stochastic gradient descent algorithms for stochastic nonconvex optimization. *arXiv preprint arXiv:1905.05920*.

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.

Wangni, J., Wang, J., Liu, J., and Zhang, T. (2018). Gradient sparsification for communication-efficient distributed optimization. *Advances in Neural Information Processing Systems*, 31.

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.

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.

Xie, C., Koyejo, S., and Gupta, I. (2019). Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance. In *International Conference on Machine Learning*, pages 6893–6901. PMLR.

Yang, Y.-R. and Li, W.-J. (2021). Basgd: Buffered asynchronous sgd for byzantine learning. In *International conference on machine learning*, pages 11751–11761. PMLR.

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.

Yu, H., Jin, R., and Yang, S. (2019). On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *International Conference on Machine Learning*, pages 7184–7193. 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*.

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 Related Work

Byzantine-robust distributed learning. A distributed learning algorithm is Byzantine-robust if its performance remains reliable in the face of Byzantine workers, which may act arbitrarily. The framework for Byzantine-robust distributed learning generally includes three main steps Guerraoui et al. (2024): (i) pre-processing of vectors submitted by workers (e.g., models or stochastic gradients), (ii) robust aggregation of these vectors, and (iii) the application of an optimization method. Existing works in this domain often vary in their treatment of one or more of these steps. For pre-processing, techniques such as Bucketing Karimireddy et al. (2022) and Nearest Neighbor Mixing (NNM) Allouah et al. (2023) have been proposed. Many robust aggregation rules have been proposed. The most commonly used ones include coordinate-wise median (CWMed), coordinate-wise trimmed mean (CWTM) Yin et al. (2018), centered clipping Karimireddy et al. (2021). Beyond the classical setting where the server outputs a single model and robustness typically relies on a sufficiently large honest fraction, list-decodable federated learning maintains a list of candidate models and guarantees that at least one model performs well even under a malicious majority Liu et al. (2025). Additionally, Byzantine attack identification strategies have demonstrated enhanced robustness in distributed computing tasks such as matrix multiplication Hong et al. (2022). Various optimization algorithms have been applied to the considered problem, including Stochastic Gradient Descent (SGD) Yang and Li (2021), Polyak momentum Karimireddy et al. (2022); El Mhamdi et al. (2021), SAGA Wu et al. (2020), and VR-MARINA Gorbunov et al. (2023).

Variance Reduction. Variance reduction is a powerful technique for accelerating the convergence of stochastic methods, particularly when a good approximation of the solution is required. The earliest variance-reduced methods were introduced by Schmidt et al. (2017); Johnson and Zhang (2013); Defazio et al. (2014). Optimal methods for variance reduction in (strongly) convex problems were later proposed by Lan and Zhou (2018); Allen-Zhu (2018); Lan et al. (2019), while methods for non-convex optimization were developed by Nguyen et al. (2017); Fang et al. (2018); Li et al. (2021). Although variance reduction has attracted significant attention Gower et al. (2020), there has been limited research on combining it with Byzantine robustness, with only a few studies addressing this area Wu et al. (2020); Zhu and Ling (2021); Karimireddy et al. (2021); Gorbunov et al. (2023).

Compressed Communications and Error Feedback. Research on distributed methods with communication compression can generally be divided into two main categories. The first focuses on methods utilizing unbiased compression operators, such as Rand$K$ sparsification (Rand$_k$) Horvóth et al. (2022), while the second explores methods that employ biased compressors, like Top$K$ sparsification (Top$_k$) Alistarh et al. (2018). A detailed summary of the most commonly used compression operators can be found in Safaryan et al. (2022); Beznosikov et al. (2023). Biased compression combined with error feedback has demonstrated strong practical performance Seide et al. (2014); Vogels et al. (2019). In the non-convex setting, which is the focus of our work, standard error feedback has been analyzed by Karimireddy et al. (2019); Beznosikov et al. (2023) and Sahu et al. (2021). However, the complexity bounds for standard error feedback typically depend on the heterogeneity parameter ζ^2 or require bounded gradients. Richtárik et al. (2021) address these challenges by introducing a novel variant of error feedback, known as EF21. This approach has since been extended in various directions by Fatkhullin et al. (2021).

B Intuition Behind Double Momentum

In this section, we show that double momentum can reduce the noise variance in a one-dimensional toy example.

We consider a simple one-dimensional stochastic gradient model

$$g_t = \mu_t + \xi_t,$$

where $\mu_t = \mathbb{E}[g_t \mid \mathcal{F}_{t-1}]$ is the (possibly time-varying) true gradient at step t, and ξ_t is the noise term with $\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0$ and $\mathrm{Var}(\xi_t \mid \mathcal{F}_{t-1}) \le \sigma^2$.

Here $\mathcal{F}_{t-1}$ denotes the sigma-field generated by all randomness up to time $t-1$. We do not need to assume that μ_t is constant or independent of the noise. The only structural assumption is the standard one for SGD-type methods: the noise sequence (ξ_t) is a martingale difference with bounded conditional variance. For the exact variance formulas below, we further consider the homoskedastic case $\mathrm{Var}(\xi_t \mid \mathcal{F}_{t-1}) = \sigma^2$.

Single momentum (conditional analysis)

For convenience of analysis, we set $v_0 = 0$; however, choosing $v_0 = g_0$ instead does not change the asymptotic variance. The single-momentum update is

$$v_{t+1} = (1-\eta)v_t + \eta g_{t+1}, \quad 0 < \eta < 1,\ v_0 = 0.$$

Unrolling the recursion yields the identity

$$v_t = \eta \sum_{k=0}^{t-1} (1-\eta)^k g_{t-k}.$$

Substituting $g_{t-k} = \mu_{t-k} + \xi_{t-k}$, we decompose

$$v_t = \underbrace{\eta \sum_{k=0}^{t-1} (1-\eta)^k \mu_{t-k}}_{\text{signal part}} + \underbrace{\eta \sum_{k=0}^{t-1} (1-\eta)^k \xi_{t-k}}_{\text{noise part}}.$$

Now fix any realization of $\mu_{1:t} = \{\mu_1, \dots, \mu_t\}$ and take conditional expectation with respect to the noise:

$$\mathbb{E}[v_t \mid \mu_{1:t}] = \eta \sum_{k=0}^{t-1} (1-\eta)^k \mu_{t-k}.$$

This shows that in the time-varying case, momentum tracks an exponentially weighted average of past true gradients, not the instantaneous μ_t. This is the usual bias of momentum methods.

For the conditional variance, define the centered noise part

$$\widetilde{v}_t := v_t - \mathbb{E}[v_t \mid \mu_{1:t}] = \eta \sum_{k=0}^{t-1} (1-\eta)^k \xi_{t-k}.$$

Using the martingale-difference property of the noise terms ξ_s and the assumption $\mathrm{Var}(\xi_s \mid \mathcal{F}_{s-1}) = \sigma^2$, we obtain

$$\mathbb{E}[\widetilde{v}_t^2 \mid \mu_{1:t}] = \eta^2 \sigma^2 \sum_{k=0}^{t-1} (1-\eta)^{2k}.$$

Letting $t \to \infty$ gives

$$\mathrm{Var}_{\mathrm{noise}}(v_\infty \mid \mu_{1:\infty}) = \eta^2 \sigma^2 \sum_{k=0}^{\infty} (1-\eta)^{2k} = \frac{\eta}{2-\eta} \sigma^2.$$

Importantly, this conditional variance formula is identical to the constant-μ case and does not depend on how μ_t is generated.

Double momentum (conditional analysis)

Double momentum maintains two EMA variables. Starting from $v_0 = u_0 = 0$, the updates are

$$v_{t+1} = (1-\eta)v_t + \eta g_{t+1},$$
$$u_{t+1} = (1-\eta)u_t + \eta v_{t+1}.$$

Unrolling the second recursion gives

$$u_t = \eta^2 \sum_{r=0}^{t-1} (r+1)(1-\eta)^r g_{t-r}.$$

Again substituting $g_{t-r} = \mu_{t-r} + \xi_{t-r}$, we obtain

$$u_t = \underbrace{\eta^2 \sum_{r=0}^{t-1} (r+1)(1-\eta)^r \mu_{t-r}}_{\text{signal part}} + \underbrace{\eta^2 \sum_{r=0}^{t-1} (r+1)(1-\eta)^r \xi_{t-r}}_{\text{noise part}}.$$

Conditioning on $\mu_{1:t}$, the centered noise component is

$$\widetilde{u}_t := u_t - \mathbb{E}[u_t \mid \mu_{1:t}] = \eta^2 \sum_{r=0}^{t-1} (r+1)(1-\eta)^r \xi_{t-r},$$

and its conditional variance is

$$\mathbb{E}[\widetilde{u}_t^2 \mid \mu_{1:t}] = \eta^4 \sigma^2 \sum_{r=0}^{t-1} (r+1)^2 (1-\eta)^{2r}.$$

Letting $t \to \infty$ and using the standard series formula $\sum_{r=0}^{\infty} (r+1)^2 b^r = \frac{1+b}{(1-b)^3}$ with $b = (1-\eta)^2$, we get

$$\mathrm{Var}_{\text{noise}}(u_\infty \mid \mu_{1:\infty}) = \sigma^2 \eta \, \frac{2 - 2\eta + \eta^2}{(2-\eta)^3}.$$

Conditional variance comparison

Therefore,

$$\mathrm{Var}_{\text{noise}}(v_\infty \mid \mu_{1:\infty}) = \frac{\eta}{2-\eta}\sigma^2, \qquad \mathrm{Var}_{\text{noise}}(u_\infty \mid \mu_{1:\infty}) = \sigma^2 \eta \, \frac{2 - 2\eta + \eta^2}{(2-\eta)^3},$$

and the ratio

$$\frac{\mathrm{Var}_{\text{noise}}(u_\infty \mid \mu_{1:\infty})}{\mathrm{Var}_{\text{noise}}(v_\infty \mid \mu_{1:\infty})} = \frac{2 - 2\eta + \eta^2}{(2-\eta)^2} \in \left[\tfrac{1}{2}, 1\right), \quad 0 < \eta < 1.$$

Thus, the conditional noise variance of the double-momentum estimator is strictly smaller than that of the single-momentum estimator, with an asymptotic reduction factor of about $1/2$ when η is small. The time variation of μ_t only affects the conditional mean $\mathbb{E}[v_t \mid \mu_{1:t}]$ and $\mathbb{E}[u_t \mid \mu_{1:t}]$, i.e., the bias, but not the above variance formulas for the noise component.

C Further Details on Robust Aggregation and Byzantine Attacks

C.1 Robust Aggregation

In Section 2, we apply robust aggregation rules satisfying Definition 2.5 after using NNM proposed by Allouah et al. (2023) (see Algorithm 2). This algorithm can enhance the robustness of aggregation rules. In particular, Allouah et al. (2023) shows that Algorithm 2 makes Robust Federated Averaging (RFA) Pillutla et al. (2022) (also known as geometric median), Coordinate-wise Median (CM) Chen et al. (2017) robust, and Coordinate-wise Trimmed Mean (CWTM) in view of the definition from Allouah et al. (2023).

Algorithm 2 NNM: Nearest Neighbor Mixing Allouah et al. (2023)

Input: number of inputs n, number of Byzantine inputs $B<n/2$, vectors $x_1, \ldots, x_n \in \mathbb{R}^d$.
for $i = 1 \ldots n$ **do**
  Sort inputs to get $(x_{i:1}, \ldots, x_{i:n})$ such that $\|x_{i:1} - x_i\| \leq \ldots \leq \|x_{i:n} - x_i\|$;
  Average the G nearest neighbors of x_i, i.e., $y_i = \frac{1}{G} \sum_{j=1}^{G} x_{i:j}$;
end for
Return: $y_1, \ldots, y_n$;

Our main goal in this section is to show that RFA $\circ$ NNM, CM $\circ$ NNM, and CWTM $\circ$ NNM satisfy Definition 2.5. Before we prove this fact, we need to introduce RFA, CM, CWTM.

Robust Federated Averaging. RFA-estimator finds a geometric median:

$$\text{RFA}(x_1, \ldots, x_n) \stackrel{\text{def}}{=} \operatorname*{argmin}_{x \in \mathbb{R}^d} \sum_{i=1}^{n} \|x - x_i\|. \tag{14}$$

The above problem has no closed-form solution. However, one can compute approximate RFA using several steps of smoothed Weiszfeld algorithm having $\mathcal{O}(n)$ computation cost of each iteration Weiszfeld (1937); Pillutla et al. (2022).

Coordinate-wise Median. CM-estimator computes a median of each component separately. That is, for the t-th coordinate, it is defined as

$$[\text{CM}(x_1, \ldots, x_n)]_t \stackrel{\text{def}}{=} \text{Median}([x_1]_t, \ldots, [x_n]_t) = \operatorname*{argmin}_{u \in \mathbb{R}} \sum_{i=1}^{n} |u - [x_i]_t|, \tag{15}$$

Coordinate-wise Trimmed Mean. The CWTM-estimator for input vectors $x_1, \ldots, x_n \in \mathbb{R}^d$ is defined as:

$$[\text{CWTM}(x_1, \ldots, x_n)]_t \stackrel{\text{def}}{=} \operatorname*{argmin}_{v \in \mathbb{R}} \sum_{i=B+1}^{n-B} ([x_i]_t - v)^2, \tag{16}$$

where $[x_i]_t$ represents the t-th coordinate values of the input vectors, sorted in non-decreasing order. CWTM finds the value v that minimizes the sum of squared differences with the middle $(n - 2B)$ values after discarding the smallest B and largest B.

C.2 Byzantine Attacks

Byzantine workers send sophisticated malicious updates to the server. In this scenario, we model an omniscient attacker Baruch et al. (2019), who knows the data of all clients. Therefore, the attacker can mimic the statistical properties of honest updates and then craft a malicious one.

- **Sign Flipping (SF)** Allen-Zhu et al. (2021): Byzantine workers compute $-c_i^{t+1}$ and send it to the server.
- **Label Flipping (LF)** Allen-Zhu et al. (2021): 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 the empirical mean $\mu_{\mathcal{G}}$ and standard deviation $\sigma_{\mathcal{G}}$ of $\{c_i^{t+1}\}_{i\in\mathcal{G}}$ and send $\mu_{\mathcal{G}} - z\sigma_{\mathcal{G}}$ 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 $-\frac{z}{G}\sum_{i\in\mathcal{G}} c_i^{t+1}$ to the server, where $z > 0$ is a constant that controls the strength of the attack.

D Extra Experiments and Experimental Details

D.1 General Setup

Our running environment has the following setup:

- CPU: AMD Ryzen Threadripper PRO 5975WX 32-Cores,
- GPU: NVIDIA GeForce RTX 4090 with CUDA version 11.7,
- PyTorch version: 2.0.1.

D.2 Datasets

We apply the proposed algorithm to two types of classification tasks to evaluate its robustness, namely binary classification and image classification. For the binary classification task, we utilize the a9a and w8a datasets from LIBSVM Chang and Lin (2011). The information of the a9a and the w8a dataset is summarized in Table 2). For the image classification task, we use the CIFAR-10 Krizhevsky et al. (2009) and FEMNIST Caldas et al. (2018) datasets. The introduction of the data set and the distribution of the data are as follows:

Table 2: Overview of the LibSVM datasets used.

Dataset	N (# of datapoints)	d (# of features)
`a9a`	32,561	123
`w8a`	49,749	300

- **CIFAR-10.** The CIFAR-10 dataset consists of 60,000 32 × 32 color images across 10 classes, with 6,000 images per class. The sampled images are evenly distributed among the clients. For CIFAR-10, we train a Residual Network with 20 layers (ResNet-20) He et al. (2016).
- **FEMNIST.** The Federated Extended MNIST (FEMNIST) dataset is a widely used benchmark for federated learning, constructed by partitioning the EMNIST dataset Cohen et al. (2017). FEMNIST consists of 805,263 images across 62 classes. For this study, we randomly sample 5% of the images from the original dataset and distribute them to clients in an independent and identically distributed (i.i.d.) manner. It is important to note that we simulate an imbalanced data distribution, where the number of training samples varies across clients. The implementation is based on LEAF Caldas et al. (2018). For FEMNIST, we train a Convolutional Neural Network (CNN) Krizhevsky et al. (2009) with two convolutional layers.

In each case, the data is distributed among $n = 20$ workers, out of which 8 are Byzantine.

D.3 Experiments Setup

For each experiment, we select the step size from the candidate set {0.5, 0.05, 0.005} and fix it throughout the training process. No learning rate warmup or decay is applied. Each experiment is repeated with three different random seeds, and we report the mean training loss or testing accuracy along with one standard error. Additionally, we adopt the same set of robust aggregation hyperparameters as outlined in Karimireddy et al. (2022), i.e.,

- **Byz-EF21-SGDM, Byz-DM21, and Byz-VR-DM21 :**the momentum parameter is set as $\eta = 0.1$,

- **Byz-VR-MARINA :** the probability to compute full gradient as $p = \min\left\{{}^{b}/_{m}, {}^{1}/_{(1+\omega)}\right\}$,
- **BR-DIANA :** the compressed difference parameter $\beta = 0.01$,
- **RFA :** the number of steps of smoothed Weisfield algorithm $T = 8$,
- **ALIE :** a small constant that controls the strength of the attack z is chosen according to Baruch et al. (2019),
- **IPM :** a small constant that controls the strength of the attack $\epsilon = 0.1$.

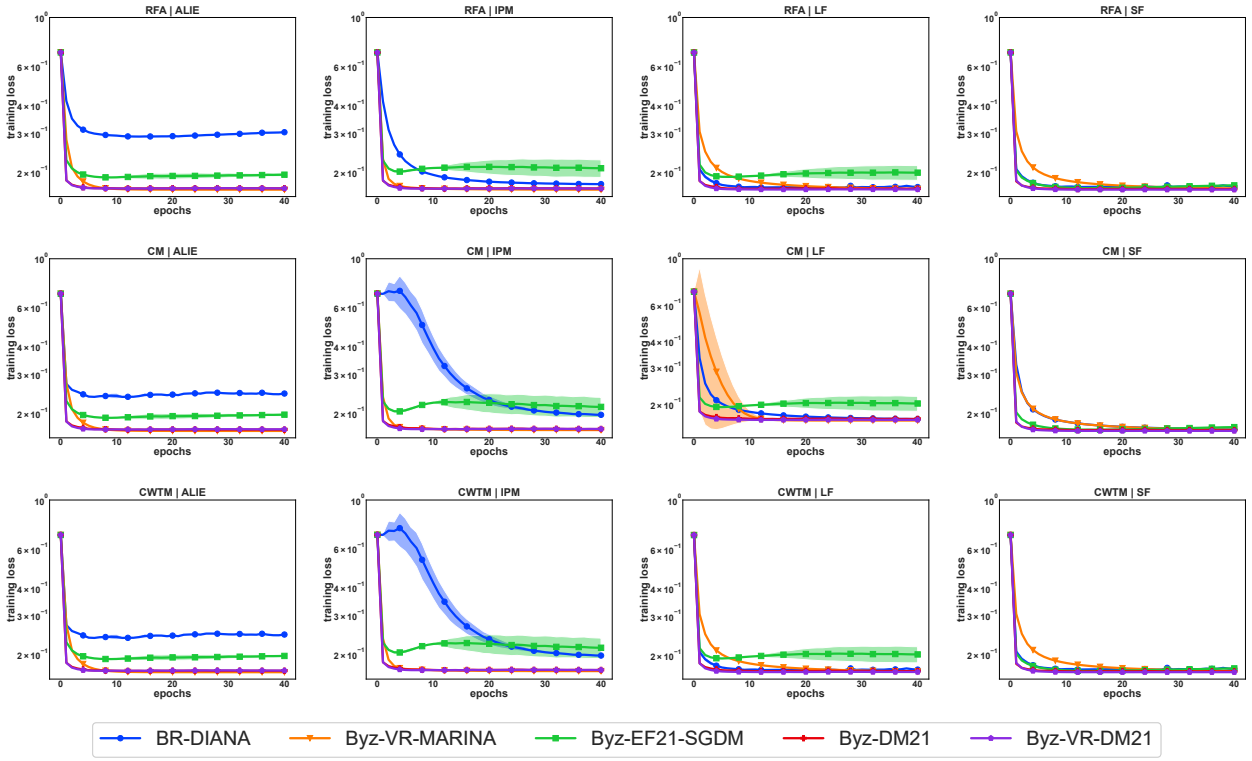

Figure 3: The training loss of RFA, CM, and CWTM aggregation rules under four attack scenarios (SF, IPM, LF, ALIE) on the w8a dataset in a heterogeneous setting. BR-DIANA and Byz-VR-MARINA use the Rand_k compressor with $k = 0.1d$, while Byz-EF21-SGDM, Byz-DM21, and Byz-VR-DM21 use the Top_k compressor with $k = 0.1d$.

D.4 Empirical Results on logistic regression

For this task, we consider solving a logistic regression problem with l_2-regularization:

$$f(x, \xi_i) = \log(1 + \exp(-b_i a_i x)) + \lambda \|x\|^2$$

where $\xi_i = (a_i, b_i) \in \mathbb{R}^d \times \{-1, 1\}$ denotes each data point, with $a_i \in \mathbb{R}^d$ and $b_i \in \{-1, 1\}$, and $\lambda > 0$ is the regularization parameter. We set $\lambda = 1/m$ for these experiments, where m is the number of samples in the local datasets. For all methods, we use a batch size of $b = 1$, and select the step-size from the following candidates: $\gamma \in \{0.5, 0.05, 0.005\}$. We use $k = 0.1d$ for both Top_k and Rand_k compressors. Finally, the number of epochs is set to 40. To ensure reproducibility, all experiments were conducted using three different random seeds. We report the mean training loss along with one standard error.

Figure 3 illustrates the training loss for five algorithms—BR-DIANA, Byz-VR-MARINA, Byz-EF21-SGDM, Byz-DM21, and Byz-VR-DM21—under a 0.4 adversarial setting on the w9a dataset in a heterogeneous setting. The results show that Byz-DM21 and Byz-VR-DM21 overall have a slight advantage, being able to converge more quickly to the optimal solution, even when there is a high proportion of Byzantine workers. The improvement in performance is attributed to the implementation of the double momentum mechanism and variance reduction techniques, which effectively reduce the impact of malicious updates.

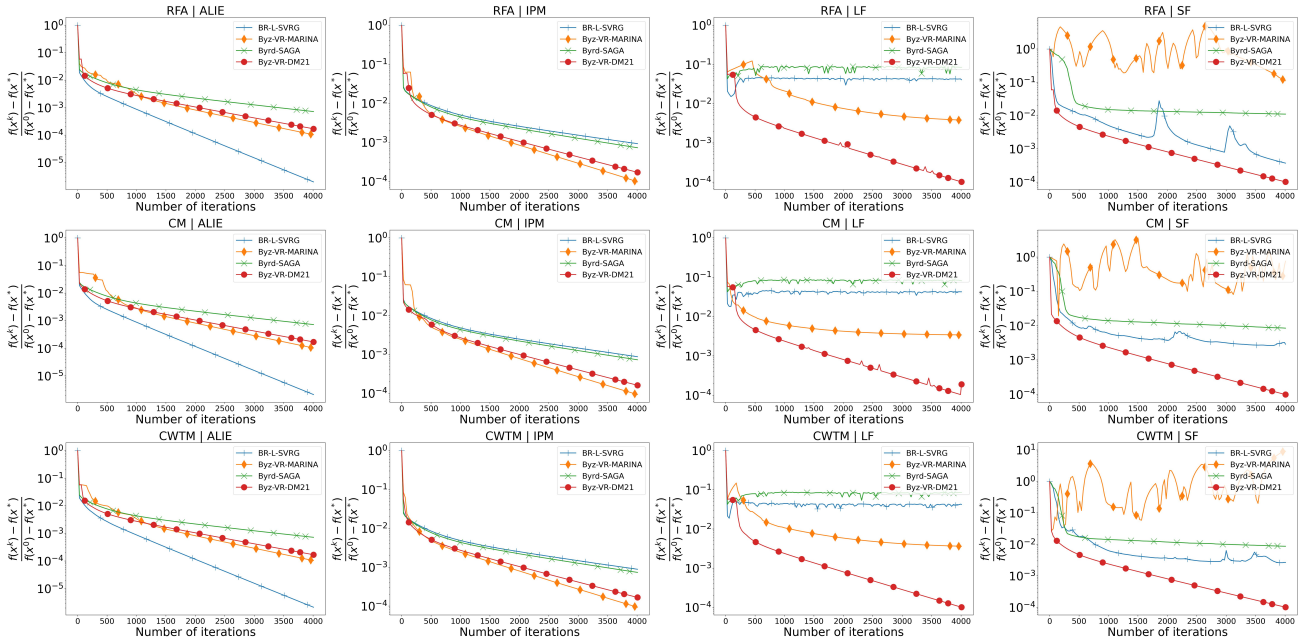

Figure 4: The relative error curve of 3 aggregation rules (RFA, CM, CWTM) under 4 attacks (ALIE, IPM, LF, SF) on the a9a dataset in a heterogeneous setting. The dataset is uniformly split over 12 honest workers with 8 Byzantine workers. BR-LSVRG, Byz-VR-MARINA, Byrd-SAGA and Byz-VR-DM21 with batchsize $b = 0.01m$ and step size $\gamma = 1/2L$.

D.5 Comparison of Variance Reduction Methods

In this experiment, we compare our Byz-VR-DM21 with several well-known variance-reduction algorithms, including BR-LSVRG Fedin and Gorbunov (2023), Byz-VR-MARINA Gorbunov et al. (2023), and Byrd-SAGA Wu et al. (2020). We evaluate these methods on the a9a dataset from the LIBSVM library, using logistic regression with l_2 regularization. Specifically, we set $l_2 = L/1000$ in our experiments, where L denotes the L-smoothness constant. The total number of workers is $n = 20$, including 8 Byzantine workers. All methods were evaluated with a step size $\gamma = 1/2L$ and a batch size of $b = 0.01m$, i.e., $b = 325$.

Figure 4 presents the relative error curve of four different algorithms—BR-LSVRG, Byz-VR-MARINA, Byrd-SAGA, and Byz-VR-DM21—under a 0.4 adversarial setting on the a9a dataset in a heterogeneous setting. The results show that, under both Sign Flipping (SF) and Label Flipping (LF) attacks, Byz-VR-DM21 consistently outperforms the other algorithms, achieving faster convergence to the optimal solution. This underscores Byz-VR-DM21's superior robustness in minimizing the impact of adversarial updates, enabling the model to stay on course toward optimal performance despite the presence of Byzantine workers. BR-LSVRG is particularly effective in the A Little Is Enough (ALIE) attacks, and Byz-VR-MARINA is particularly effective in the Inner Product Manipulation (IPM) attacks. In all cases, Byz-VR-DM21 demonstrates convergence to very high accuracy, while none of the algorithms dominate across all the attacks. In addition, Byz-VR-DM21 consistently achieves a functional suboptimality of at least 10^{-4}, representing reasonably good accuracy. This experiment demonstrates that Byz-VR-DM21 can converge to high accuracy even with small or moderate batch sizes.

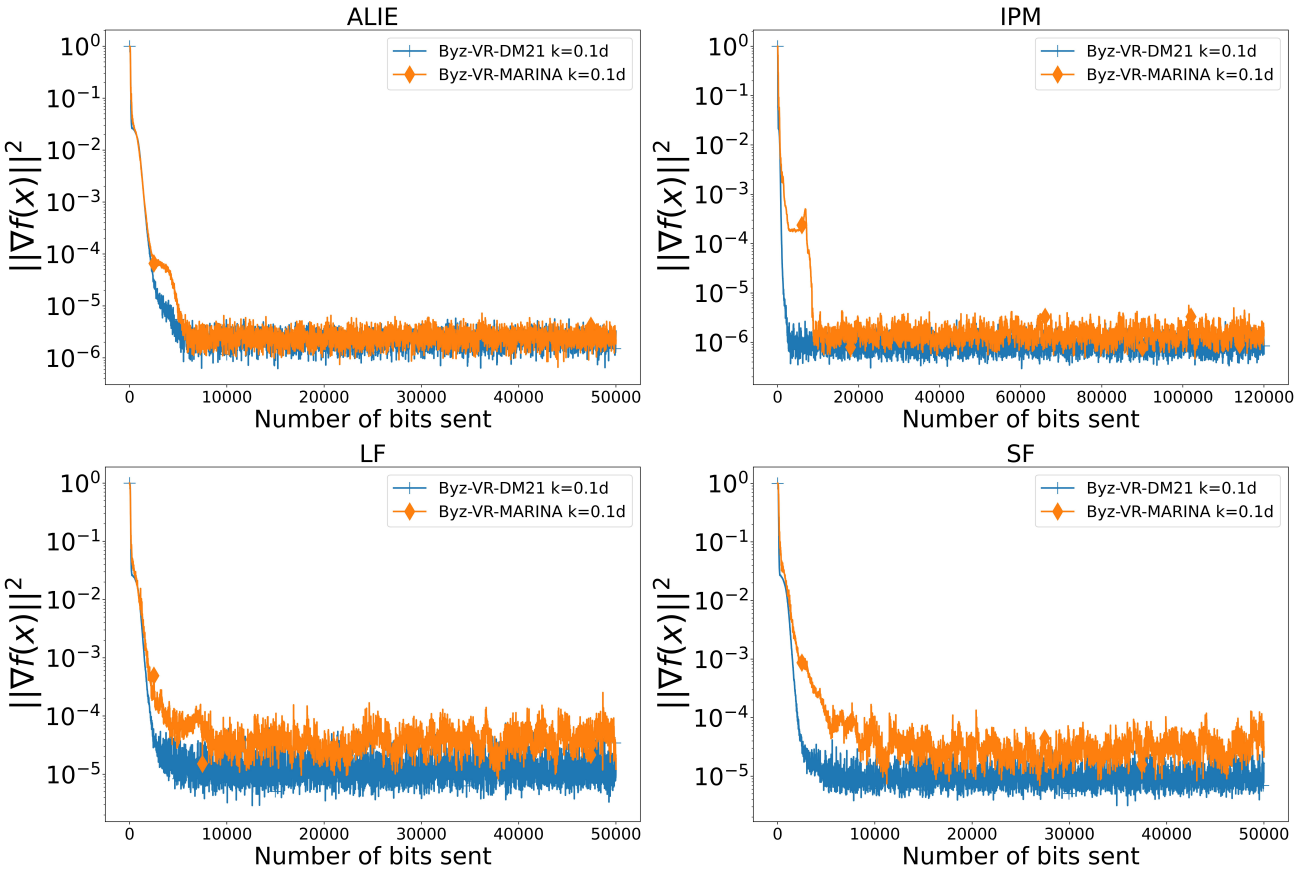

Figure 5: The communication complexity comparison under 4 attacks (ALIE, IPM, LF, SF) on the a9a dataset in a heterogeneous setting. Byz-VR-MARINA uses the Rand_k compressor, while Byz-VR-DM21 uses the Top_k compressor, with $k = 0.1d$, batch size $b = 1$, and step size $\gamma = 1/2L$.

D.6 Error Feedback Experiments

We next compare the empirical performance of Byz-VR-DM21 and Byz-VR-MARINA on the a9a dataset in a heterogeneous setting. The total number of workers is $n = 20$, including 2 Byzantine workers. All methods were evaluated with a step size of $\gamma = 1/2L$ and a batch size of $b = 1$. We use $k = 0.1d$ for both the Top_k and Rand_k compressors.

The experiments unveil promising potential for error feedback of communication improvement. As shown in Figures 5, Byz-VR-DM21 converges slightly faster than Byz-VR-MARINA, before both reach a point of stagnation.

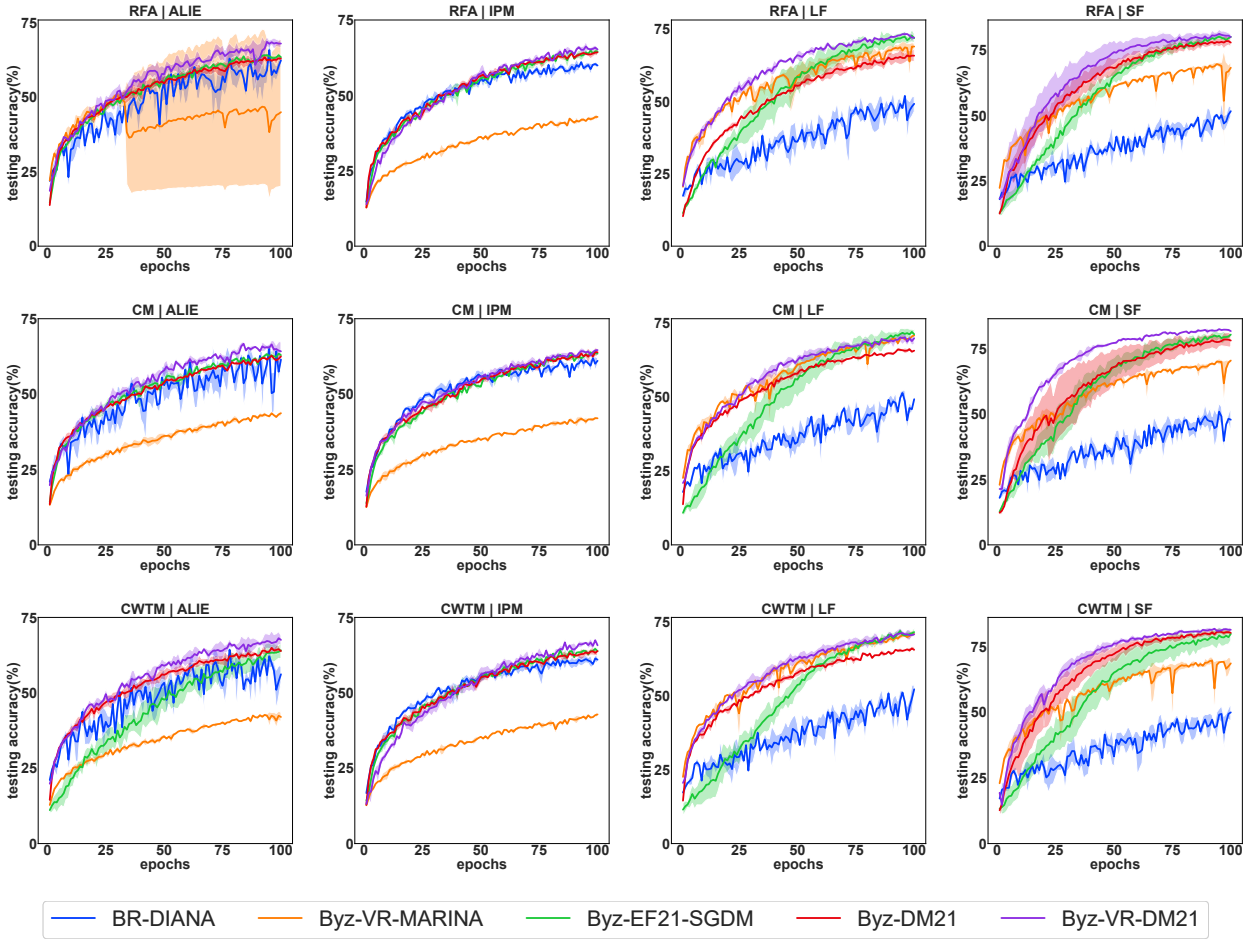

Figure 6: The testing accuracy (%) of 3 aggregation rules (RFA, CM, CWTM) under 4 attacks (ALIE, IPM, LF, SF) on the CIFAR-10 dataset. BR-DIANA and Byz-VR-MARINA use the Rand_k compressor, while Byz-EF21-SGDM, Byz-DM21, and Byz-VR-DM21 use the Top_k compressor, where $k = 0.1d$.

D.7 Empirical Results on CIFAR-10

We evaluate our algorithms on an image classification task using the CIFAR-10 dataset Krizhevsky et al. (2009) with the ResNet-20 deep neural network He et al. (2016). For all methods, we use a batch size of $b = 64$ and select the step size from the following candidates: $\gamma \in \{0.5, 0.05, 0.005\}$. We use $k = 0.1d$ for both Top_k and Rand_k compressors. The training process is carried out over 100 epochs, iterating over 5000 iterations. To ensure reproducibility, all experiments were conducted using three different random seeds. We report the mean testing accuracy along with one standard error.

Figure 6 presents the testing accuracy of five different algorithms—BR-DIANA, Byz-VR-MARINA, Byz-EF21-SGDM, Byz-DM21, and Byz-VR-DM21—under a 0.4 adversarial setting on the CIFAR-10 dataset. The results show that Byz-VR-DM21 demonstrates a slight advantage over the other algorithms in all the attack scenarios considered, even in the case of a high proportion of Byzantine workers.

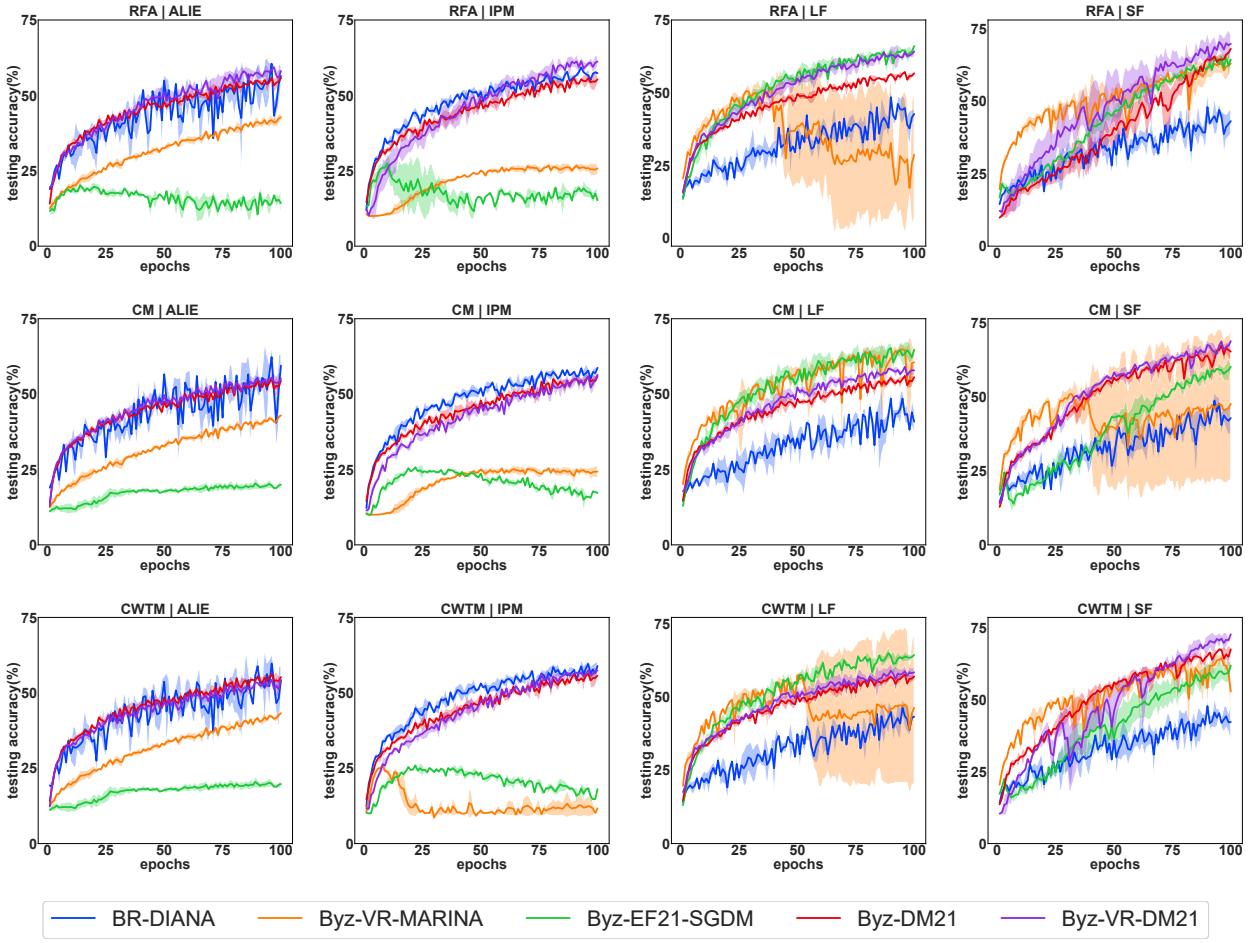

Figure 7: The testing accuracy (%) of 3 aggregation rules (RFA, CM, CWTM) under 4 attacks (ALIE, IPM, LF, SF) on the CIFAR-10 dataset with a heterogeneity regime of $a = 0.25$ (high). BR-DIANA and Byz-VR-MARINA use the Rand_k compressor, while Byz-EF21-SGDM, Byz-DM21, and Byz-VR-DM21 use the Top_k compressor, where $k = 0.1d$.

Figure 7 presents the testing accuracy of five different algorithms—BR-DIANA, Byz-VR-MARINA, Byz-EF21-SGDM, Byz-DM21, and Byz-VR-DM21—under a 0.4 adversarial setting on the CIFAR-10 dataset with a heterogeneity regimes of $a = 0.25$ (high). The results show that Byz-VR-DM21 demonstrates a slight advantage over the other algorithms in all the attack scenarios considered, even in the case of a high proportion of Byzantine workers and high heterogeneity.

In Table 3, we present the performance of five algorithms—BR-DIANA, Byz-VR-MARINA, Byz-EF21-SGDM, Byz-DM21, and Byz-VR-DM21—on the CIFAR-10 dataset. For each algorithm and under every attack, we highlight in bold the algorithm that results in the highest accuracy for the considered scenario.

In Table 4, we present the performance of five algorithms—BR-DIANA, Byz-VR-MARINA, Byz-EF21-SGDM, Byz-DM21, and Byz-VR-DM21—on the CIFAR-10 dataset with a heterogeneity regimes of $a = 0.25$ (high). For each algorithm and under every attack, we highlight in bold the algorithm that results in the highest accuracy for the considered scenario.

Aggregation	Method	ALIE	IPM	LF	SF	N.A.	Worst Case
RFA+NNM	BR-DIANA	65.86 ± 02.03	61.03 ± 00.83	51.96 ± 02.66	51.62 ± 02.07	67.64 ± 01.63	51.62 ± 02.07
	Byz-VR-MARINA	50.76 ± 24.86	42.86 ± 00.79	69.03 ± 01.17	69.63 ± 02.44	71.07 ± 00.18	42.86 ± 00.79
	Byz-EF21-SGDM	64.14 ± 00.33	64.90 ± 01.35	72.75 ± 01.55	80.12 ± 00.87	80.94 ± 01.12	64.14 ± 00.33
	Byz-DM21	63.47 ± 01.41	64.26 ± 01.00	65.98 ± 02.28	78.42 ± 01.63	80.10 ± 02.14	63.47 ± 01.41
	Byz-VR-DM21	**68.63 ± 01.37**	**66.20 ± 00.64**	**73.34 ± 00.51**	**81.19 ± 01.21**	**82.33 ± 00.87**	66.20 ± 00.64
CM+NNM	BR-DIANA	64.99 ± 02.05	61.93 ± 01.02	51.27 ± 01.72	50.91 ± 04.73	67.95 ± 02.05	50.91 ± 04.73
	Byz-VR-MARINA	43.66 ± 01.24	42.02 ± 00.28	70.71 ± 00.36	70.46 ± 00.37	70.17 ± 01.66	42.02 ± 00.28
	Byz-EF21-SGDM	63.42 ± 00.64	63.88 ± 00.64	**71.89 ± 01.54**	80.36 ± 01.06	80.98 ± 01.02	63.42 ± 00.64
	Byz-DM21	62.77 ± 00.45	63.96 ± 01.18	66.04 ± 00.49	78.50 ± 02.53	80.73 ± 00.44	62.77 ± 00.45
	Byz-VR-DM21	**66.70 ± 03.19**	**64.60 ± 00.62**	70.08 ± 01.99	**82.46 ± 00.59**	**82.66 ± 00.31**	64.60 ± 00.62
CWTM+NNM	BR-DIANA	64.29 ± 02.73	61.41 ± 00.97	52.09 ± 01.49	49.97 ± 03.31	68.35 ± 02.42	49.97 ± 03.31
	Byz-VR-MARINA	42.68 ± 01.53	42.82 ± 01.05	70.71 ± 00.67	69.75 ± 01.93	70.57 ± 01.10	42.68 ± 01.53
	Byz-EF21-SGDM	64.10 ± 02.96	64.56 ± 00.32	**71.78 ± 01.75**	79.89 ± 02.51	79.89 ± 02.51	64.10 ± 02.96
	Byz-DM21	64.90 ± 00.29	63.86 ± 01.57	65.89 ± 01.14	80.71 ± 01.59	80.93 ± 01.68	63.86 ± 01.57
	Byz-VR-DM21	**68.11 ± 01.46**	**67.41 ± 00.45**	71.04 ± 00.77	**81.87 ± 00.88**	**83.08 ± 00.49**	67.41 ± 00.45

Table 3: Maximum testing accuracy (%) across $T = 5000$ learning steps on the CIFAR-10 dataset, under four Byzantine attacks strategies. There are $B = 8$ Byzantine workers among $n = 20$. 'N.A.' denotes the case with no Byzantine attackers. In each of the three horizontal blocks and under each attack, the best accuracy is highlighted in **bold**. Additionally, for every method, we report the worst-case accuracy across attacks.

Aggregation	Method	ALIE	IPM	LF	SF	N.A.	Worst Case
RFA+NNM	BR-DIANA	**60.52 ± 03.29**	59.28 ± 01.08	48.47 ± 01.74	48.00 ± 04.43	62.25 ± 01.63	48.00 ± 04.43
	Byz-VR-MARINA	42.77 ± 01.34	26.76 ± 01.97	51.72 ± 21.35	63.06 ± 02.42	64.15 ± 02.18	26.76 ± 01.97
	Byz-EF21-SGDM	20.23 ± 04.62	27.41 ± 04.01	**66.05 ± 01.52**	65.25 ± 01.54	72.62 ± 00.38	20.23 ± 04.62
	Byz-DM21	55.80 ± 00.61	55.64 ± 01.30	57.13 ± 00.72	68.13 ± 01.00	75.77 ± 00.85	55.64 ± 01.30
	Byz-VR-DM21	59.36 ± 00.92	**61.36 ± 01.67**	64.15 ± 01.77	**70.49 ± 02.92**	**78.96 ± 00.74**	59.36 ± 00.92
CM+NNM	BR-DIANA	**62.34 ± 03.85**	**58.71 ± 00.96**	48.54 ± 03.82	48.81 ± 05.03	63.24 ± 01.86	47.04 ± 01.86
	Byz-VR-MARINA	42.94 ± 00.66	25.37 ± 01.86	64.84 ± 04.01	50.97 ± 25.65	63.04 ± 02.61	25.37 ± 01.86
	Byz-EF21-SGDM	20.36 ± 01.84	25.70 ± 00.39	**65.31 ± 02.07**	60.27 ± 00.83	71.57 ± 00.97	20.36 ± 01.84
	Byz-DM21	55.23 ± 00.42	55.91 ± 00.44	56.08 ± 00.57	67.49 ± 02.84	76.59 ± 01.72	55.23 ± 00.42
	Byz-VR-DM21	55.44 ± 02.34	56.39 ± 01.93	59.05 ± 01.93	**68.96 ± 01.14**	**79.57 ± 01.89**	55.44 ± 02.34
CWTM+NNM	BR-DIANA	**59.71 ± 05.10**	**59.58 ± 01.54**	46.39 ± 00.73	47.83 ± 07.10	61.00 ± 01.25	46.39 ± 00.73
	Byz-VR-MARINA	43.28 ± 00.23	25.93 ± 02.38	57.55 ± 25.73	64.58 ± 12.93	64.74 ± 12.54	25.93 ± 02.38
	Byz-EF21-SGDM	20.46 ± 01.78	25.91 ± 00.55	**64.98 ± 00.21**	61.98 ± 02.01	72.75 ± 00.77	20.46 ± 01.78
	Byz-DM21	56.27 ± 00.27	56.34 ± 00.96	57.71 ± 01.20	67.55 ± 00.59	72.30 ± 05.23	56.27 ± 00.27
	Byz-VR-DM21	54.31 ± 00.32	57.41 ± 00.86	59.06 ± 02.12	**72.78 ± 02.59**	**79.23 ± 00.71**	54.31 ± 00.32

Table 4: Maximum testing accuracy (%) across $T = 5000$ learning steps on the CIFAR-10 dataset with a heterogeneity regimes of $a = 0.25$ (high), under four Byzantine attacks strategies. There are $B = 8$ Byzantine workers among $n = 20$. 'N.A.' denotes the case with no Byzantine attackers. In each of the three horizontal blocks and under each attack, the best accuracy is highlighted in **bold**. Additionally, For every method, we report the worst-case accuracy across attacks.

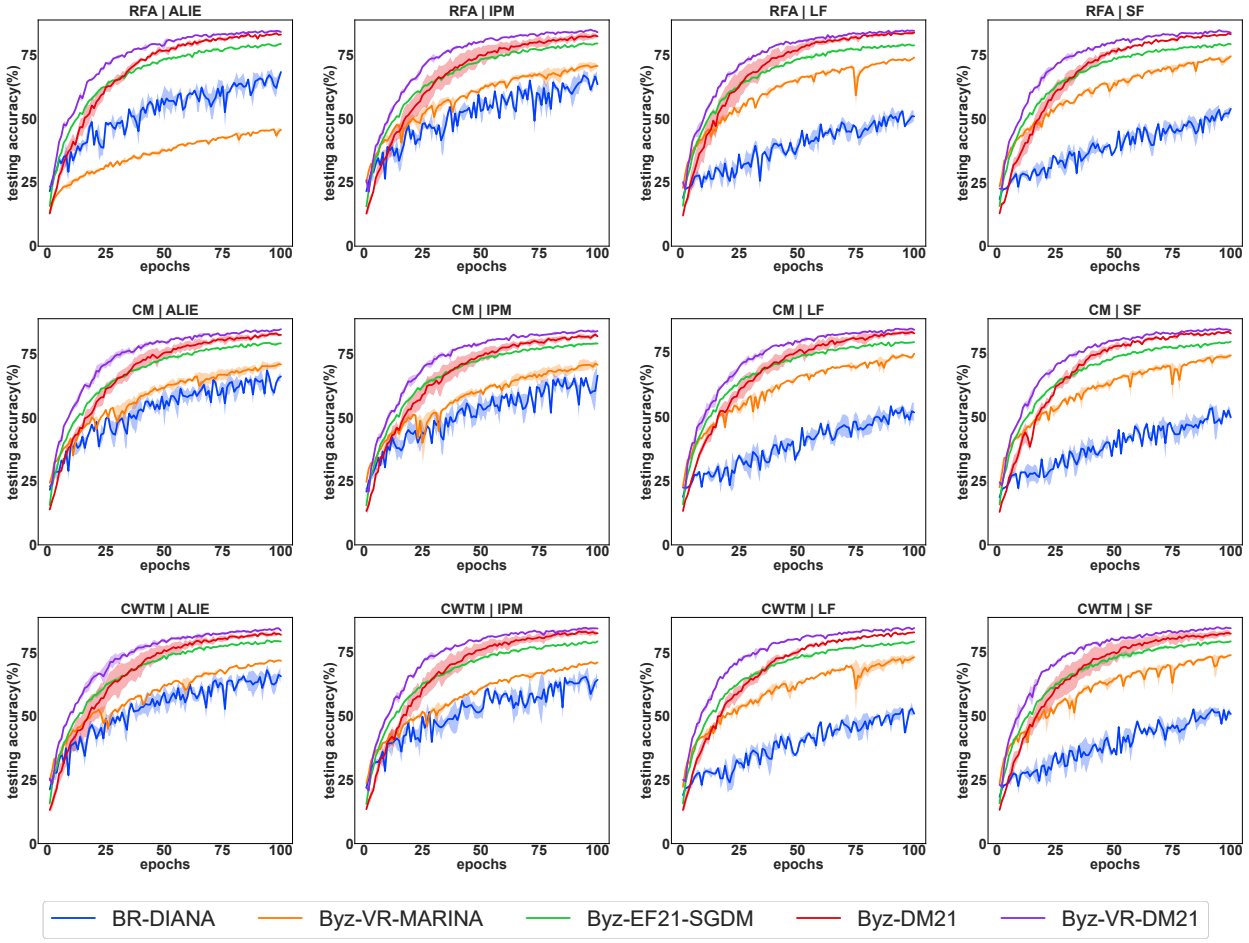

Figure 8: The testing accuracy (%) of 3 aggregation rules (RFA, CM, CWTM) under 4 attacks (ALIE, IPM, LF, SF) on the CIFAR-10 dataset. The dataset is uniformly split among 20 workers, including 1 Byzantine worker. BR-DIANA and Byz-VR-MARINA use the Rand_k compressor, while Byz-EF21-SGDM, Byz-DM21, and Byz-VR-DM21 use the Top_k compressor, where $k = 0.1d$.

D.8 Byzantine Ratio Experiments on CIFAR-10

We evaluate validation accuracy under different Byzantine worker ratios. The results include three configurations: 19 good workers with 1 Byzantine worker(Figure 8), 18 good workers with 2 Byzantine workers(Figure 9), and 16 good workers with 4 Byzantine workers(Figure 10). For all methods, we use a batch size of $b = 64$ and select the step size from the following candidates: $\gamma \in \{0.5, 0.05, 0.005\}$. We use $k = 0.1d$ for both Top_k and Rand_k compressors. The training process is carried out over 100 epochs, iterating over 5000 iterations. To ensure reproducibility, all experiments were conducted using three different random seeds. We report the mean testing accuracy along with one standard error.

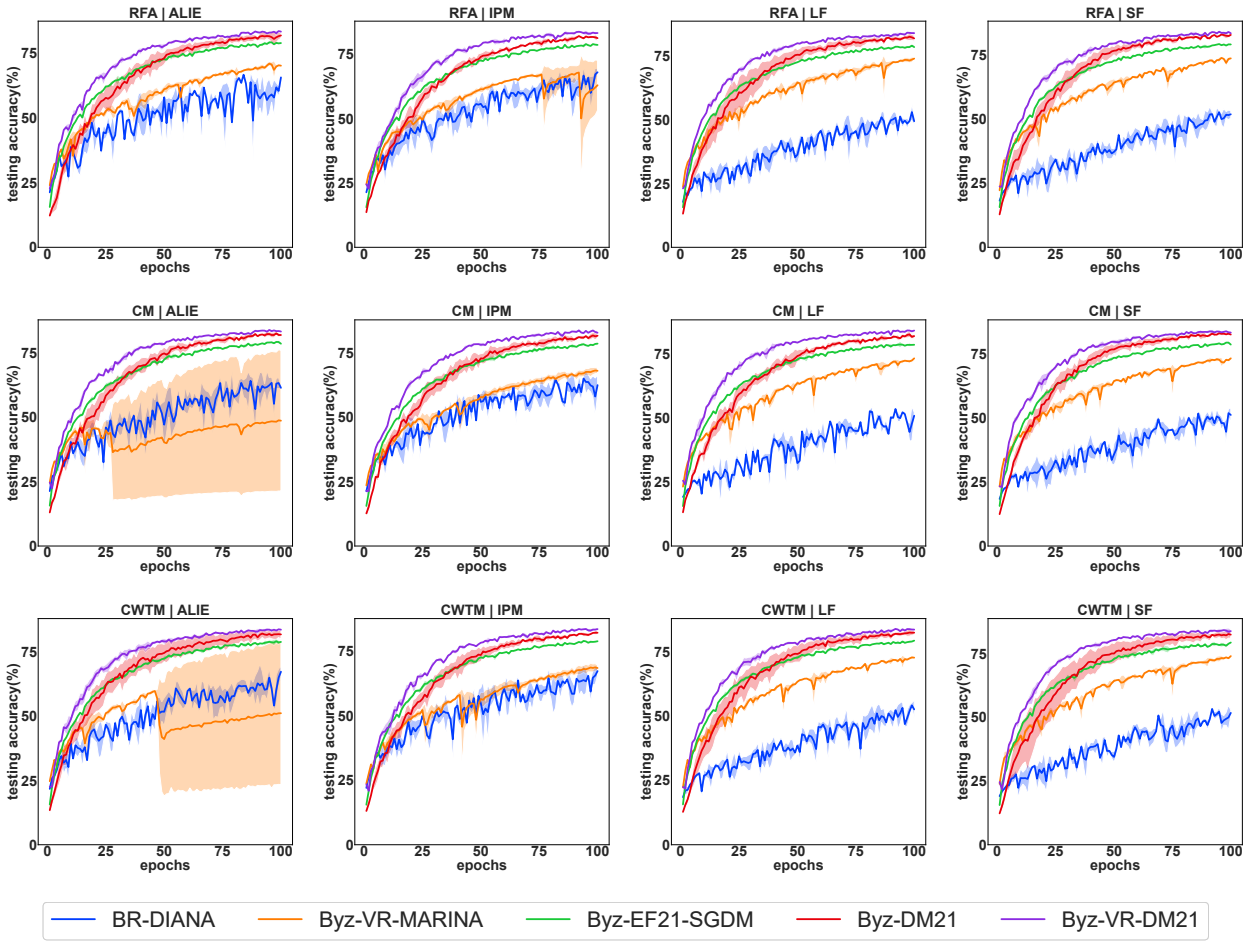

Figure 9: The testing accuracy (%) of 3 aggregation rules (RFA, CM, CWTM) under 4 attacks (ALIE, IPM, LF, SF) on the CIFAR-10 dataset. The dataset is uniformly split among 20 workers, including 2 Byzantine workers. BR-DIANA and Byz-VR-MARINA use the Rand_k compressor, while Byz-EF21-SGDM, Byz-DM21, and Byz-VR-DM21 use the Top_k compressor, where $k = 0.1d$.

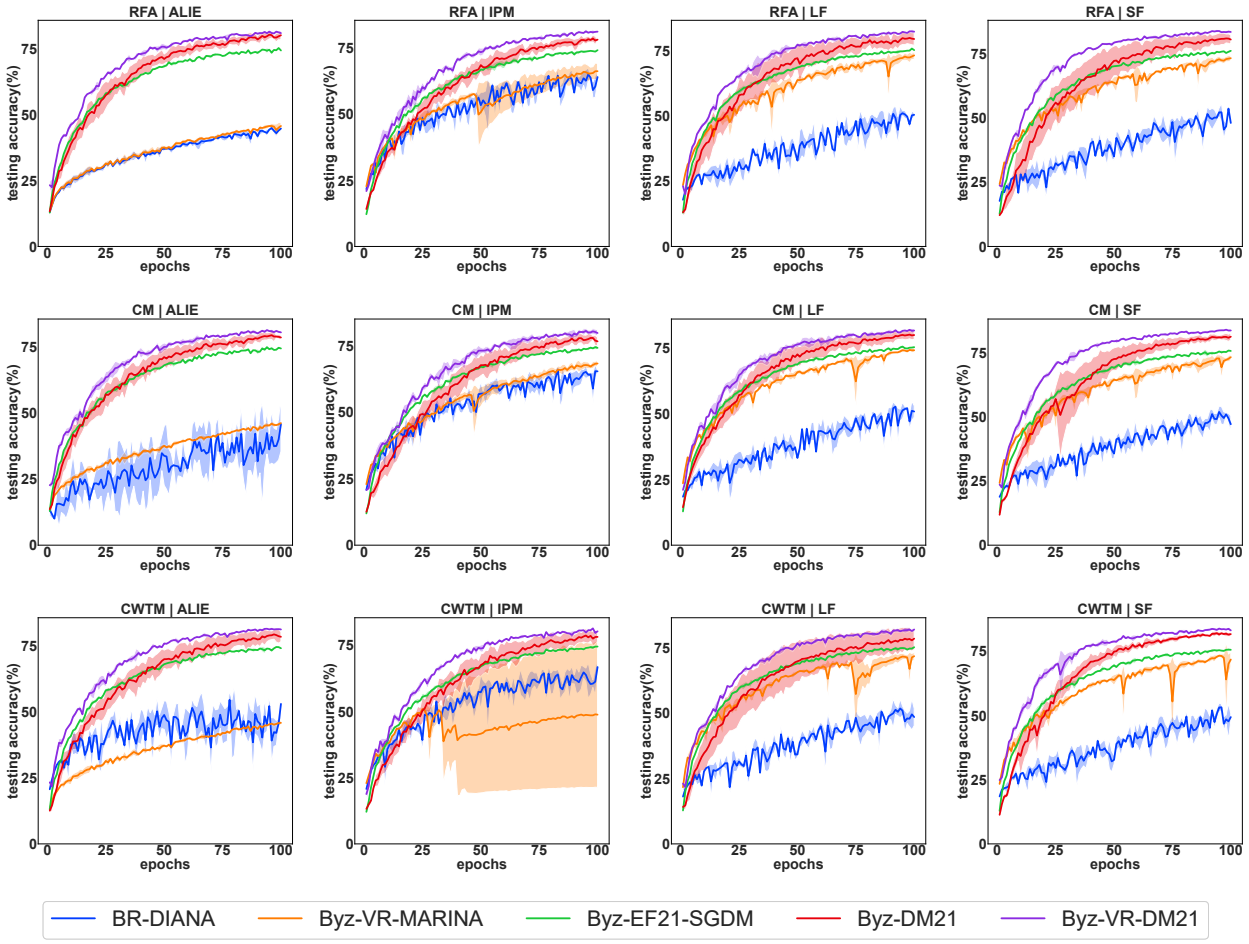

Figure 10: The testing accuracy (%) of 3 aggregation rules (RFA, CM, CWTM) under 4 attacks (ALIE, IPM, LF, SF) on the CIFAR-10 dataset. The dataset is uniformly split among 20 workers, including 4 Byzantine workers. BR-DIANA and Byz-VR-MARINA use the Rand_k compressor, while Byz-EF21-SGDM, Byz-DM21, and Byz-VR-DM21 use the Top_k compressor, where $k = 0.1d$.

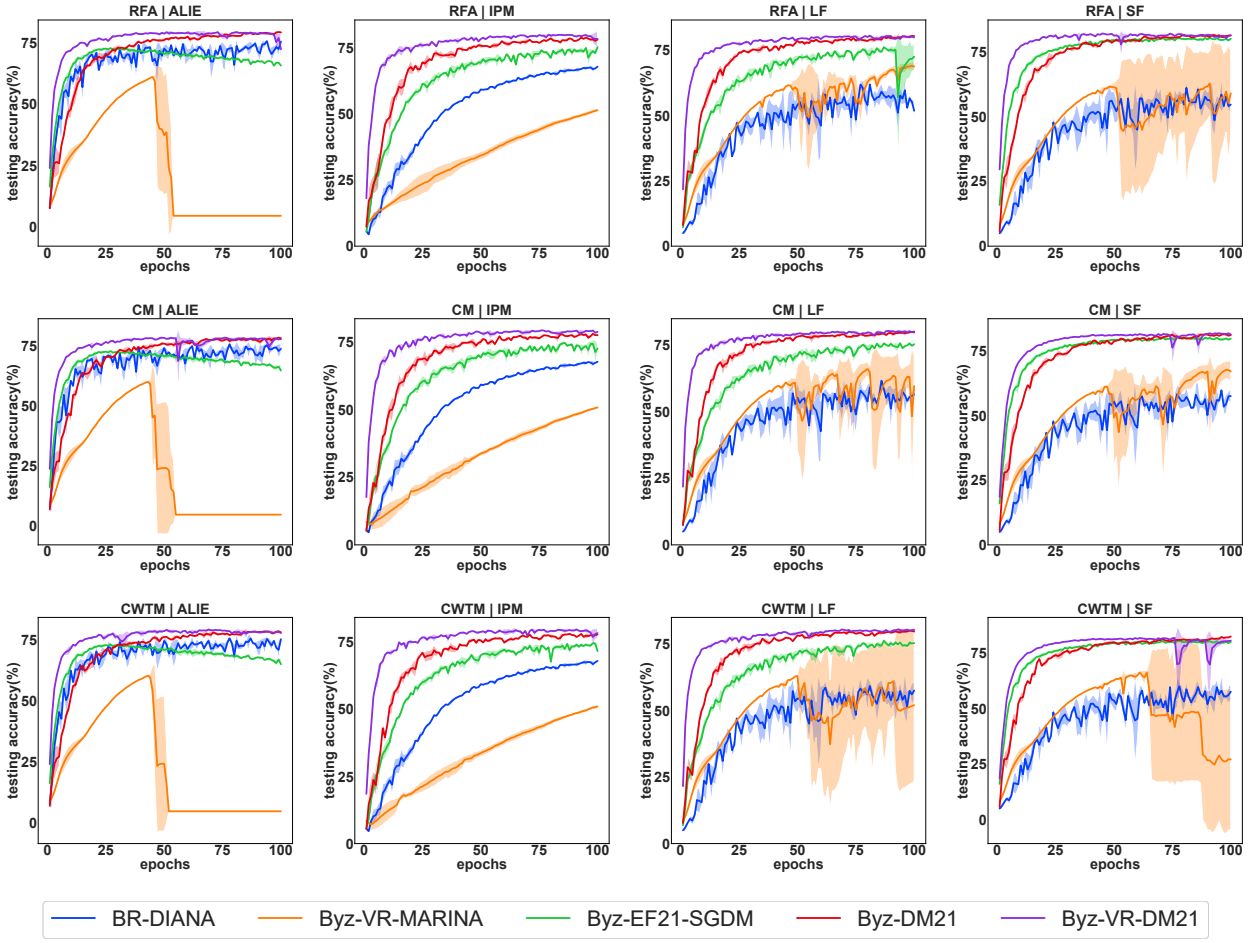

Figure 11: The testing accuracy (%) of 3 aggregation rules (RFA, CM, CWTM) under 4 attacks (ALIE, IPM, LF, SF) on the FEMNIST dataset. BR-DIANA and Byz-VR-MARINA use the Rand_k compressor, while Byz-EF21-SGDM, Byz-DM21, and Byz-VR-DM21 use the Top_k compressor, where $k = 0.1d$.

D.9 Empirical Results on FEMNIST

We evaluate our algorithms on an image classification task using the FEMNIST dataset Caldas et al. (2018) with a convolutional neural network (CNN) Krizhevsky et al. (2009). For all methods, we use a batch size of $b = 32$ and select the step size from the following candidates :$\gamma \in \{0.5, 0.05, 0.005\}$. We use $k = 0.1d$ for both Top_k and Rand_k compressors. The training process is carried out over 100 epochs, iterating over 5000 iterations. To ensure reproducibility, all experiments were conducted using three different random seeds. We report the mean testing accuracy along with one standard error.

Figure 11 presents the testing accuracy of five different algorithms—BR-DIANA, Byz-VR-MARINA, Byz-EF21-SGDM, Byz-DM21, and Byz-VR-DM21—under a 0.4 adversarial setting on the FEMNIST dataset. The results show that Byz-DM21 and Byz-VR-DM21 demonstrate a slight advantage over the other algorithms in all attack scenarios considered, achieving higher accuracy during convergence, even with a high proportion of Byzantine workers. This highlights the benefits of the double momentum and variance reduction approach.

Aggregation	Method	ALIE	IPM	LF	SF	N.A.	Worst Case
RFA+NNM	BR-DIANA	75.74 ± 00.78	67.94 ± 00.60	61.81 ± 04.07	61.06 ± 02.34	65.16 ± 00.29	61.81 ± 04.07
	Byz-VR-MARINA	61.18 ± 00.00	51.45 ± 00.83	68.91 ± 02.00	62.77 ± 16.24	70.74 ± 01.35	51.45 ± 00.83
	Byz-EF21-SGDM	72.94 ± 01.07	75.37 ± 00.49	76.07 ± 05.02	80.23 ± 00.40	80.32 ± 00.94	72.94 ± 01.07
	Byz-DM21	**79.45 ± 00.34**	79.01 ± 00.59	80.35 ± 00.14	81.49 ± 00.44	81.85 ± 00.60	79.01 ± 00.59
	Byz-VR-DM21	79.40 ± 07.33	**80.05 ± 03.27**	**80.50 ± 01.12**	**81.72 ± 00.15**	**82.02 ± 00.41**	79.40 ± 07.33
CM+NNM	BR-DIANA	75.71 ± 01.71	67.75 ± 00.67	61.61 ± 02.19	59.87 ± 02.81	73.41 ± 01.28	59.87 ± 02.81
	Byz-VR-MARINA	60.05 ± 00.00	50.85 ± 00.54	66.04 ± 11.14	67.79 ± 04.05	70.79 ± 04.79	50.85 ± 00.54
	Byz-EF21-SGDM	72.95 ± 03.18	74.71 ± 01.83	75.66 ± 00.23	80.37 ± 00.34	80.45 ± 00.15	72.95 ± 03.18
	Byz-DM21	78.52 ± 00.36	78.07 ± 00.40	79.90 ± 00.52	81.60 ± 00.59	81.65 ± 00.42	78.07 ± 00.40
	Byz-VR-DM21	**78.53 ± 00.54**	**79.37 ± 00.36**	**80.29 ± 00.42**	**81.61 ± 00.51**	**81.98 ± 00.89**	78.53 ± 00.54
CWTM+NNM	BR-DIANA	69.48 ± 01.39	67.94 ± 00.61	59.11 ± 03.31	59.96 ± 03.74	72.85 ± 02.88	59.11 ± 03.31
	Byz-VR-MARINA	60.18 ± 00.00	50.97 ± 00.79	62.90 ± 28.64	66.20 ± 31.62	69.30 ± 06.43	50.97 ± 00.79
	Byz-EF21-SGDM	72.94 ± 03.06	74.54 ± 02.51	76.12 ± 00.09	80.39 ± 00.45	80.50 ± 01.96	72.94 ± 03.06
	Byz-DM21	78.35 ± 01.13	78.00 ± 00.55	79.89 ± 00.80	81.49 ± 00.45	**82.39 ± 00.13**	78.00 ± 00.55
	Byz-VR-DM21	**78.99 ± 00.35**	**79.61 ± 01.93**	**80.35 ± 00.34**	**81.67 ± 00.66**	81.86 ± 00.89	78.99 ± 00.35

Table 5: Maximum testing accuracy (%) across $T = 5000$ learning steps on the FEMNIST dataset, under four Byzantine attack strategies. There are $B = 8$ Byzantine workers among $n = 20$. 'N.A.' denotes the case with no Byzantine attackers. In each of the three horizontal blocks and under each attack, the best accuracy is highlighted in **bold**. Additionally, for every method, we report the worst-case accuracy across attacks.

In Table 5, we present the performance of five algorithms—BR-DIANA, Byz-VR-MARINA, Byz-EF21-SGDM, Byz-DM21, and Byz-VR-DM21—on the FEMNIST dataset. For each algorithm and under every attack, we highlight in bold the algorithm that results in the highest accuracy for the considered scenario.

D.10 Additional Experiments

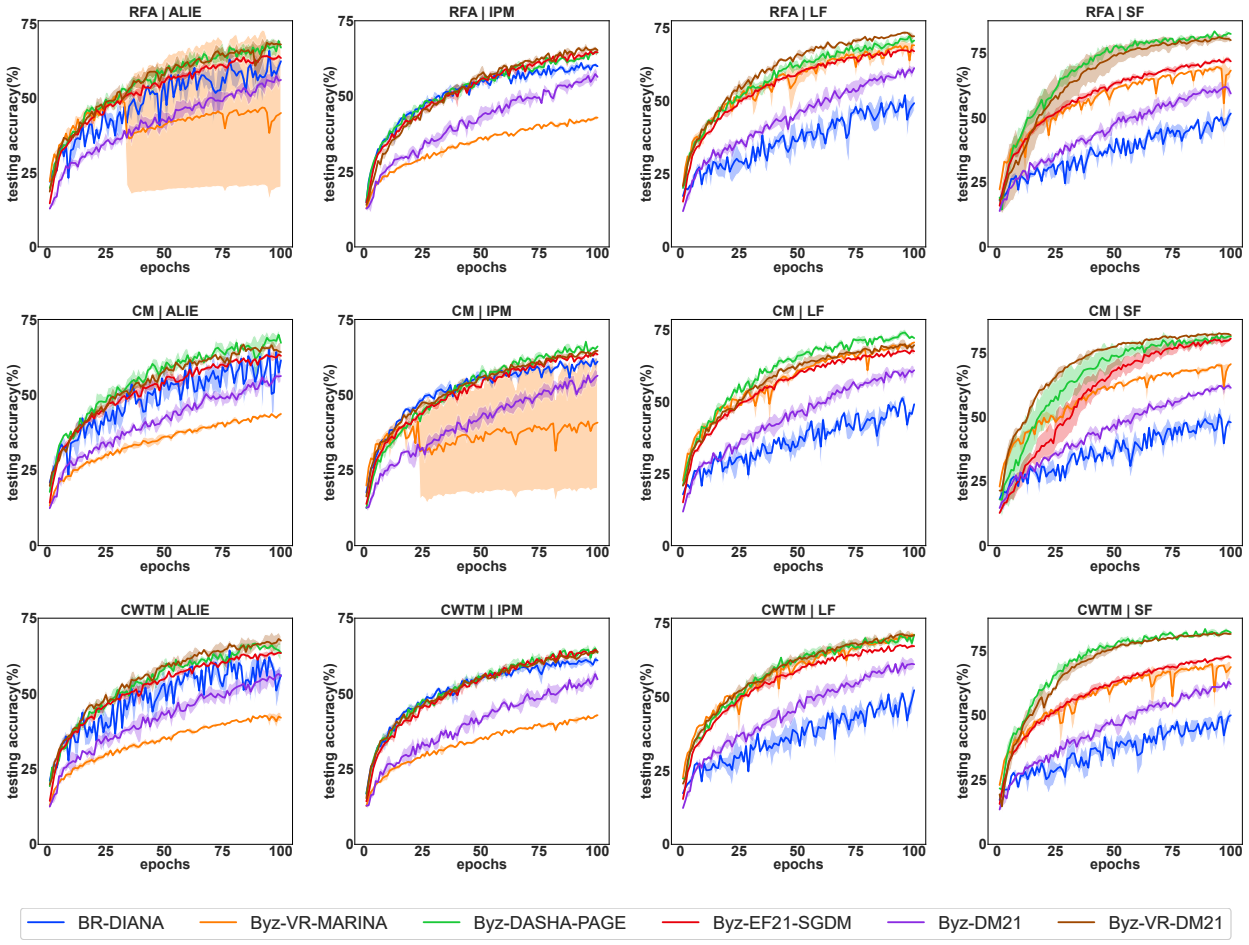

Figure 12: The testing accuracy (%) of 3 aggregation rules (RFA, CM, CWTM) under 4 attacks (ALIE, IPM, LF, SF) on the CIFAR-10 dataset. BR-DIANA, Byz-VR-MARINA, and Byz-DASHA-PAGE use the Rand_k compressor, while Byz-EF21-SGDM, Byz-DM21, and Byz-VR-DM21 use the Top_k compressor, where $k = 0.1d$.

Figure 12 presents the testing accuracy of six different algorithms—BR-DIANA, Byz-VR-MARINA, Byz-DASHA-PAGE,Byz-EF21-SGDM, Byz-DM21, and Byz-VR-DM21—under a 0.4 adversarial setting on the CIFAR-10 dataset. The results show that Byz-VR-DM21 consistently outperforms the other baselines across all considered attack scenarios, even when the proportion of Byzantine workers is high; the only exception is Byz-DASHA-PAGE. Under our current CIFAR-10 setup and hyperparameters, Byz-DASHA-PAGE is very competitive empirically and in some cases slightly better than Byz-DM21, and close to Byz-VR-DM21 in terms of test accuracy. However, this comes at a non-negligible algorithmic cost: Byz-DASHA-PAGE requires computing a full local gradient with probability at each iteration, while our methods (Byz-DM21 and Byz-VR-DM21) never rely on full-gradient evaluations and only use stochastic gradients. In federated or large-scale distributed settings, full gradients can be significantly more expensive than stochastic ones, especially when each client holds a large dataset. For this reason, we view Byz-VR-DM21 as offering a more favorable trade-off between robustness, convergence guarantees, and per-iteration computational cost, even when its final accuracy is comparable to that of Byz-DASHA-PAGE.

E Useful Facts

For all $a, b \in \mathbb{R}^d$ and $\alpha > 0, \rho \in (0, 1]$ the following relations hold:

$$\|a+b\|^2 \leq (1+\rho)\|a\|^2 + (1+\rho^{-1})\|b\|^2, \tag{17}$$

$$\|a+b+c\|^2 \leq 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2, \tag{18}$$

$$\|a+b+c+d\|^2 \leq 4\|a\|^2 + 4\|b\|^2 + 4\|c\|^2 + 4\|d\|^2. \tag{19}$$

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

$$\mathbb{E}\Big[\|X-c\|^2\Big] = \mathbb{E}\Big[\|X-\mathbb{E}[X]\|^2\Big] + \|\mathbb{E}[X]-c\|^2. \tag{20}$$

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

F Missing Proofs of Byz-DM21 for General Non-Convex Functions

Let us state the following lemma used in the analysis of our methods.

F.1 Supporting Lemmas

We prepare the following lemmas to facilitate the proof of Theorem 3.1.

Lemma F.1 (Descent lemma from Li et al. (2021)). *Given an L-smooth function $f(x)$. For the update $x^{(t+1)} = x^{(t)} - \gamma g^{(t)}$, there holds*

$$f(x^{(t+1)}) \leq f(x^{(t)}) - \frac{\gamma}{2}\|\nabla f(x^{(t)})\|^2 - (\frac{1}{2\gamma} - \frac{L}{2})\|x^{(t+1)} - x^{(t)}\|^2 + \frac{\gamma}{2}\|g^{(t)} - \nabla f(x^{(t)})\|^2. \tag{21}$$

Lemma F.2 (Robust aggregation error). *Suppose that Assumption 2.3 holds. Then, for all $t \geq 0$ the iterates generated by* Byz-DM21 *in Algorithm 1 satisfy the following condition:*

$$\begin{aligned}
&\|g^{(t)} - \overline{g}^{(t)}\|^2 \\
&\leq \frac{\kappa}{G}\sum_{i\in\mathcal{G}}\|g_i^{(t)} - \overline{g}^{(t)}\|^2 \\
&\leq \frac{\kappa}{2G^2}\sum_{i,j\in\mathcal{G}}\|g_i^{(t)} - g_j^{(t)}\|^2 \\
&\leq \frac{8\kappa(G-1)}{G^2}\sum_{i\in\mathcal{G}}\left(\|C_i^{(t)}\|^2 + \|P_i^{(t)}\|^2 + \|M_i^{(t)}\|^2\right) + \frac{8\kappa(G-1)}{G}\zeta^2,
\end{aligned} \tag{22}$$

where $\overline{g}^{(t)} = G^{-1}\sum_{i\in\mathcal{G}} g_i$, $C_i^{(t)} \overset{\text{def}}{=} g_i^{(t)} - u_i^{(t)}$, $P_i^{(t)} \overset{\text{def}}{=} u_i^{(t)} - v_i^{(t)}$, $M_i^{(t)} \overset{\text{def}}{=} v_i^{(t)} - \nabla f_i(x^{(t)})$.

Proof. Define $H_i^{(t)} \overset{\text{def}}{=} \nabla f_i(x^{(t)}) - \nabla f(x^{(t)})$. We consider

$$\begin{aligned}
&\sum_{i,j\in\mathcal{G}}\|g_i^{(t)} - g_j^{(t)}\|^2 \\
&= \sum_{i,j\in\mathcal{G}, i\neq j}\|C_i^{(t)} + P_i^{(t)} + M_i^{(t)} + H_i^{(t)} - C_j^{(t)} - P_j^{(t)} - M_j^{(t)} - H_j^{(t)}\|^2 \\
&\leq 8\sum_{i,j\in\mathcal{G}, i\neq j}\left(\|C_i^{(t)}\|^2 + \|P_i^{(t)}\|^2 + \|M_i^{(t)}\|^2 + \|H_i^{(t)}\|^2 + \|C_j^{(t)}\|^2 + \|P_j^{(t)}\|^2 + \|M_j^{(t)}\|^2 + \|H_j^{(t)}\|^2\right) \\
&= 16(G-1)\sum_{i\in\mathcal{G}}\left(\|C_i^{(t)}\|^2 + \|P_i^{(t)}\|^2 + \|M_i^{(t)}\|^2 + \|H_i^{(t)}\|^2\right) \\
&\leq 16(G-1)\sum_{i\in\mathcal{G}}\left(\|C_i^{(t)}\|^2 + \|P_i^{(t)}\|^2 + \|M_i^{(t)}\|^2\right) + 16(G-1)G\zeta^2.
\end{aligned}$$

Since

$$\sum_{i\in\mathcal{G}}\|g_i^{(t)} - \overline{g}^{(t)}\|^2 = \frac{1}{2G}\sum_{i,j\in\mathcal{G}}\|g_i^{(t)} - g_j^{(t)}\|^2,$$

and the aggregation mechanism is (B,κ)-robust[4], we have

$$\begin{aligned}
&\|g^{(t)} - \overline{g}^{(t)}\|^2 \\
&\leq \frac{\kappa}{G}\sum_{i\in\mathcal{G}}\|g_i^{(t)} - \overline{g}^{(t)}\|^2 \\
&\leq \frac{\kappa}{2G^2}\sum_{i,j\in\mathcal{G}}\|g_i^{(t)} - g_j^{(t)}\|^2 \\
&\leq \frac{8\kappa(G-1)}{G^2}\sum_{i\in\mathcal{G}}\left(\|C_i^{(t)}\|^2 + \|P_i^{(t)}\|^2 + \|M_i^{(t)}\|^2\right) + \frac{8\kappa(G-1)}{G}\zeta^2.
\end{aligned}$$

[4]Note that the B represents the number of Byzantine workers in this context, and $n - B$ has been replaced with G.

□

Lemma F.3 (Accumulated compression error)**.** *Let Assumption 2.1 and 2.4 be satisfied, and suppose $\mathcal{C}$ is a contractive compressor. For every $i = 1, \ldots, G$, let the sequences $\{v_i^{(t)}\}_{t\geq 0}$, $\{u_i^{(t)}\}_{t\geq 0}$, and $\{g_i^{(t)}\}_{t\geq 0}$ be updated via*

$$
\begin{aligned}
g_i^{(t)} &= g_i^{(t-1)} + \mathcal{C}(u_i^{(t)} - g_i^{(t-1)}) \\
u_i^{(t)} &= u_i^{(t-1)} + \eta(v_i^{(t)} - u_i^{(t-1)}) \\
v_i^{(t)} &= v_i^{(t-1)} + \eta(\nabla f_i(x^{(t)}, \xi_i^{(t)}) - v_i^{(t-1)}).
\end{aligned}
$$

Then for all $t \geq 0$ the iterates generated by Byz-DM21 *in Algorithm 1 satisfy*

$$
\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E}\left[\|C_i^{(t)}\|^2\right] &= \sum_{t=0}^{T-1} \mathbb{E}\left[\|g_i^{(t)} - u_i^{(t)}\|^2\right] \\
&\leq \frac{12\eta^4}{\alpha^2} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f_i(x^{(t)}) - v_i^{(t)}\|^2\right] + \frac{2\eta^4 T\sigma^2}{\alpha} \\
&\quad + \frac{12\eta^4 L_i^2}{\alpha^2} \sum_{t=0}^{T-1} \mathbb{E}\left[\|x^{(t+1)} - x^{(t)}\|^2\right] + \frac{12\eta^2}{\alpha^2} \mathbb{E}\left[\|u_i^{(t)} - v_i^{(t)}\|^2\right].
\end{aligned} \tag{23}
$$

Proof. By the update rules of $g_i^{(t)}$, $u_i^{(t)}$ and $v_i^{(t)}$, we derive

$$
\begin{aligned}
&\mathbb{E}\left[\|g_i^{(t)} - u_i^{(t)}\|^2\right] \\
&= \mathbb{E}\left[\|g_i^{(t-1)} - u_i^{(t)} + \mathcal{C}(u_i^{(t)} - g_i^{(t-1)})\|^2\right] \\
&= \mathbb{E}\left[\mathbb{E}_{\mathcal{C}}\left[\|u_i^{(t)} - g_i^{(t-1)} - \mathcal{C}(u_i^{(t)} - g_i^{(t-1)})\|^2\right]\right] \\
&\overset{(i)}{\leq} (1-\alpha)\mathbb{E}\left[\|u_i^{(t)} - g_i^{(t-1)}\|^2\right] \\
&= (1-\alpha)\mathbb{E}\left[\|u_i^{(t-1)} - g_i^{(t-1)} + \eta(v_i^{(t)} - u_i^{(t-1)})\|^2\right] \\
&= (1-\alpha)\mathbb{E}\Big[\|u_i^{(t-1)} - g_i^{(t-1)} + \eta\Big(v_i^{(t-1)} - u_i^{(t-1)} + \eta(\nabla f_i(x^{(t)}, \xi_i^{(t)}) - v_i^{(t-1)})\Big)\|^2\Big] \\
&= (1-\alpha)\mathbb{E}\Big[\|u_i^{(t-1)} - g_i^{(t-1)} + \eta\Big(v_i^{(t-1)} - u_i^{(t-1)} + \eta(\nabla f_i(x^{(t)}) - v_i^{(t-1)}) \\
&\qquad + \eta(\nabla f_i(x^{(t)}, \xi_i^{(t)}) - \nabla f_i(x^{(t)})\Big)\|^2\Big] \\
&= (1-\alpha)\mathbb{E}\Big[\mathbb{E}_{\xi_i^{(t)}}\Big[\|u_i^{(t-1)} - g_i^{(t-1)} + \eta(v_i^{(t-1)} - u_i^{(t-1)}) + \eta^2(\nabla f_i(x^{(t)}) - v_i^{(t-1)}) \\
&\qquad + \eta^2(\nabla f_i(x^{(t)}, \xi_i^{(t)}) - \nabla f_i(x^{(t)})\|^2\Big]\Big] \\
&= (1-\alpha)\mathbb{E}\Big[\|u_i^{(t-1)} - g_i^{(t-1)} + \eta(v_i^{(t-1)} - u_i^{(t-1)}) + \eta^2(\nabla f_i(x^{(t)}) - v_i^{(t-1)})\|^2\Big] \\
&\qquad + (1-\alpha)\eta^4\mathbb{E}\Big[\|\nabla f_i(x^{(t)}, \xi_i^{(t)}) - \nabla f_i(x^{(t)})\|^2\Big],
\end{aligned}
$$

where (i) refers to the contractive property, as defined in Definition 2.7, and then we use inequality (17) to obtain

$$
\begin{aligned}
&\overset{(i)}{\leq} (1-\alpha)(1+\rho)\mathbb{E}\left[\|u_i^{(t-1)} - g_i^{(t-1)}\|^2\right] + \eta^4\sigma^2 \\
&\quad + (1-\alpha)(1+\rho^{-1})\mathbb{E}\left[\|\eta(v_i^{(t-1)} - u_i^{(t-1)}) + \eta^2(\nabla f_i(x^{(t)}) - v_i^{(t-1)})\|^2\right] \\
&= (1-\alpha)(1+\rho)\mathbb{E}\left[\|u_i^{(t-1)} - g_i^{(t-1)}\|^2\right] + (1-\alpha)(1+\rho^{-1})\mathbb{E}\Big[\|\eta(v_i^{(t-1)} - u_i^{(t-1)}) \\
&\quad + \eta^2(\nabla f_i(x^{(t-1)}) - v_i^{(t-1)}) + \eta^2(\nabla f_i(x^{(t)}) - \nabla f_i(x^{(t-1)}))\|^2\Big] + \eta^4\sigma^2 \\
&\leq (1-\alpha)(1+\rho)\mathbb{E}\left[\|u_i^{(t-1)} - g_i^{(t-1)}\|^2\right] + \eta^4\sigma^2 + 3(1-\alpha)(1+\rho^{-1})\eta^2\mathbb{E}\left[\|v_i^{(t-1)} - u_i^{(t-1)}\|^2\right] \\
&\quad + 3(1-\alpha)(1+\rho^{-1})\eta^4\mathbb{E}\left[\|\nabla f_i(x^{(t-1)}) - v_i^{(t-1)}\|^2\right] \\
&\quad + 3(1-\alpha)(1+\rho^{-1})\eta^4\mathbb{E}\left[\|\nabla f_i(x^{(t)}) - \nabla f_i(x^{(t-1)})\|^2\right] \\
&\overset{(ii)}{\leq} (1-\alpha)(1+\rho)\mathbb{E}\left[\|u_i^{(t-1)} - g_i^{(t-1)}\|^2\right] + \eta^4\sigma^2 + 3(1-\alpha)(1+\rho^{-1})\eta^2\mathbb{E}\left[\|v_i^{(t-1)} - u_i^{(t-1)}\|^2\right] \\
&\quad + 3(1-\alpha)(1+\rho^{-1})\eta^4\mathbb{E}\left[\|\nabla f_i(x^{(t-1)}) - v_i^{(t-1)}\|^2\right] \\
&\quad + 3(1-\alpha)(1+\rho^{-1})\eta^4 L_i^2\mathbb{E}\left[\|x^{(t)} - x^{(t-1)}\|^2\right],
\end{aligned}
$$

where (i) refers to the bounded variance assumption, as stated in Definition 2.4, and (ii) leverages the smoothness property of $f_i(\cdot)$. By setting $\rho = \alpha/2$, we obtain the following result.

$$
(1-\alpha)(1+\frac{\alpha}{2}) = 1 - \frac{\alpha}{2} - \frac{\alpha^2}{2} \leq 1 - \frac{\alpha}{2},
$$

and

$$
(1-\alpha)(1+\rho^{-1}) = \frac{2}{\alpha} - \alpha - 1 \leq \frac{2}{\alpha}.
$$

Therefore we attain

$$
\begin{aligned}
&\mathbb{E}\left[\|g_i^{(t)} - u_i^{(t)}\|^2\right] \\
&= \mathbb{E}\left[\|g_i^{(t-1)} - u_i^{(t)} + \mathcal{C}(u_i^{(t)} - g_i^{(t-1)})\|^2\right] \\
&\leq \left(1 - \frac{\alpha}{2}\right)\mathbb{E}\left[\|u_i^{(t-1)} - g_i^{(t-1)}\|^2\right] + \frac{6\eta^4}{\alpha}\mathbb{E}\left[\|\nabla f_i(x^{(t-1)}) - v_i^{(t-1)}\|^2\right] + \eta^4\sigma^2 \\
&\quad + \frac{6\eta^4 L_i^2}{\alpha}\mathbb{E}\left[\|x^{(t)} - x^{(t-1)}\|^2\right] + \frac{6\eta^2}{\alpha}\mathbb{E}\left[\|u_i^{(t-1)} - v_i^{(t-1)}\|^2\right].
\end{aligned}
$$

Summing up the above inequality from $t = 0$ to $t = T-1$ leads to

$$
\begin{aligned}
\sum_{t=0}^{T-1}\mathbb{E}\left[\|g_i^{(t)} - u_i^{(t)}\|^2\right] &\leq \frac{12\eta^4}{\alpha^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f_i(x^{(t)}) - v_i^{(t)}\|^2\right] + \frac{2\eta^4 T\sigma^2}{\alpha} \\
&\quad + \frac{12\eta^4 L_i^2}{\alpha^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|x^{(t+1)} - x^{(t)}\|^2\right] + \frac{12\eta^2}{\alpha^2}\mathbb{E}\left[\|u_i^{(t)} - v_i^{(t)}\|^2\right].
\end{aligned}
$$

□

Lemma F.4 (Accumulated second momentum deviation)**.** *Let Assumption 2.1 and 2.4 be satisfied, and suppose $0 < \eta \leq 1$. For every $i = 1, \ldots, G$, let the sequences $\{v_i^{(t)}\}_{t\geq 0}$ and $\{u_i^{(t)}\}_{t\geq 0}$ be updated via*

$$
\begin{aligned}
v_i^{(t)} &= (1-\eta)v_i^{(t-1)} + \eta\nabla f_i(x^{(t)}, \xi_i^{(t)}) \\
u_i^{(t)} &= u_i^{(t-1)} + \eta(v_i^{(t)} - u_i^{(t-1)}).
\end{aligned}
$$

Then for all $t \geq 0$ the iterates generated by Byz-DM21 *in Algorithm 1 satisfy*

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{E}\left[\|P_i^{(t)}\|^2\right] &= \sum_{t=0}^{T-1}\mathbb{E}\left[\|u_i^{(t)}-v_i^{(t)}\|^2\right]\\ &\leq 6\sum_{t=0}^{T-1}\mathbb{E}\Big[\|v_i^{(t)}-\nabla f_i(x^{(t)}\|^2\Big] + 6L_i^2\sum_{t=0}^{T-1}\mathbb{E}\Big[\|x^{(t)}-x^{(t-1)}\|^2\Big] + \eta T\sigma^2.\end{aligned} \tag{24}$$

Proof. By the update rule of $u_i^{(t)}$ and $v_i^{(t)}$, we have

$$\begin{aligned}&\mathbb{E}\left[\|u_i^{(t)}-v_i^{(t)}\|^2\right]\\ &= \mathbb{E}\left[\|u_i^{(t-1)}-v_i^{(t)}+\eta(v_i^{(t)}-u_i^{(t-1)})\|^2\right]\\ &= (1-\eta)^2\mathbb{E}\left[\|u_i^{(t-1)}-v_i^{(t)}\|^2\right]\\ &= (1-\eta)^2\mathbb{E}\Big[\|(1-\eta)v_i^{(t-1)}+\eta\nabla f_i(x^{(t)},\xi_i^{(t)})-u_i^{(t-1)}\|^2\Big]\\ &= (1-\eta)^2\mathbb{E}\Big[\|(v_i^{(t-1)}-u_i^{(t-1)})+\eta(\nabla f_i(x^{(t)},\xi_i^{(t)})-v_i^{(t-1)})\|^2\Big]\\ &= (1-\eta)^2\mathbb{E}\Big[\|(u_i^{(t-1)}-v_i^{(t-1)})+\eta(v_i^{(t-1)}-\nabla f_i(x^{(t)}))+\eta(\nabla f_i(x^{(t)})-\nabla f_i(x^{(t)},\xi_i^{(t)}))\|^2\Big]\\ &= (1-\eta)^2\mathbb{E}\Big[\mathbb{E}_{\xi_i^{(t)}}\Big[\|u_i^{(t-1)}-v_i^{(t-1)}+\eta(v_i^{(t-1)}-\nabla f_i(x^{(t)}))+\eta(\nabla f_i(x^{(t)})-\nabla f_i(x^{(t)},\xi_i^{(t)}))\|^2\Big]\Big]\\ &= (1-\eta)^2\Big(\mathbb{E}\Big[\|u_i^{(t-1)}-v_i^{(t-1)}+\eta(v_i^{(t-1)}-\nabla f_i(x^{(t)}))\|^2\Big]+\eta^2\mathbb{E}\Big[\|\nabla f_i(x^{(t)},\xi_i^{(t)})-\nabla f_i(x^{(t)})\|^2\Big]\Big)\\ &\overset{(i)}{\leq} (1-\eta)^2\mathbb{E}\Big[\|u_i^{(t-1)}-v_i^{(t-1)}+\eta(v_i^{(t-1)}-\nabla f_i(x^{(t)}))\|^2\Big]+\eta^2\sigma^2\\ &\overset{(ii)}{\leq} (1-\eta)^2(1+\rho)\mathbb{E}\Big[\|u_i^{(t-1)}-v_i^{(t-1)}\|^2\Big]+\eta^2(1+\rho^{-1})\mathbb{E}\Big[\|v_i^{(t-1)}-\nabla f_i(x^{(t)})\|^2\Big]+\eta^2\sigma^2,\end{aligned}$$

where (i) utilizes Assumption 2.4, (ii) holds by inequality (17). Setting $\rho=\eta/2$, and because of $\eta\in(0,1]$, we obtain

$$(1-\eta)^2(1+\frac{\eta}{2}) = 1-\frac{3\eta}{2}+\frac{\eta^3}{2}\leq 1-\eta,$$

and

$$\eta^2(1+\frac{2}{\eta}) = \eta^2+2\eta\leq 3\eta\,.$$

There holds

$$\begin{aligned}&\mathbb{E}\left[\|u_i^{(t)}-v_i^{(t)}\|^2\right]\\ &\leq (1-\eta)\mathbb{E}\Big[\|u_i^{(t-1)}-v_i^{(t-1)}\|^2\Big]+3\eta\mathbb{E}\Big[\|v_i^{(t-1)}-\nabla f_i(x^{(t)})\|^2\Big]+\eta^2\sigma^2\\ &\overset{(i)}{\leq} (1-\eta)\mathbb{E}\Big[\|u_i^{(t-1)}-v_i^{(t-1)}\|^2\Big]+6\eta\mathbb{E}\Big[\|v_i^{(t-1)}-\nabla f_i(x^{(t-1)})\|^2\Big]\\ &\quad+6\eta\mathbb{E}\Big[\|\nabla f_i(x^{(t)})-\nabla f_i(x^{(t-1)})\|^2\Big]+\eta^2\sigma^2\\ &\overset{(ii)}{\leq} (1-\eta)\mathbb{E}\Big[\|u_i^{(t-1)}-v_i^{(t-1)}\|^2\Big]+6\eta\mathbb{E}\Big[\|v_i^{(t-1)}-\nabla f_i(x^{(t-1)})\|^2\Big]\\ &\quad+6\eta L_i^2\mathbb{E}\Big[\|x^{(t)}-x^{(t-1)}\|^2\Big]+\eta^2\sigma^2,\end{aligned}$$

where (i) uses inequality (17), and (ii) uses smoothness of $f_i(\cdot)$. Summing up the above inequality from $t=0$ to

$t = T-1$ yields

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{E}\left[\|P_i^{(t)}\|^2\right] &\le 6\sum_{t=0}^{T-1}\mathbb{E}\Big[\|v_i^{(t)}-\nabla f_i(x^{(t)}\|^2\Big] + 6L_i^2\sum_{t=0}^{T-1}\mathbb{E}\Big[\|x^{(t)}-x^{(t-1)}\|^2\Big]\\ &\quad + \frac{1}{\eta}\sum_{t=0}^{T-1}\mathbb{E}\left[\|P_i^{(0)}\|^2\right] + \eta T\sigma^2\\ &\le 6\sum_{t=0}^{T-1}\mathbb{E}\Big[\|v_i^{(t)}-\nabla f_i(x^{(t)}\|^2\Big] + 6L_i^2\sum_{t=0}^{T-1}\mathbb{E}\Big[\|x^{(t)}-x^{(t-1)}\|^2\Big] + \eta T\sigma^2,\end{aligned}$$

where $P_i^{(0)} = u_i^{(0)} - v_i^{(0)}$, because of $u_i^{(0)} = v_i^{(0)}$, $P_i^{(0)} = 0$. □

Lemma F.5 (Accumulated momentum deviation)**.** *Let Assumption 2.1 and 2.4 be satisfied, and suppose* $0 < \eta \le 1$. *For every* $i = 1, \dots, G$, *let the sequences* $\{v_i^{(t)}\}_{t\ge 0}$ *be updated via*

$$v_i^{(t)} = (1-\eta)v_i^{(t-1)} + \eta\nabla f_i(x^{(t)}, \xi_i^{(t)}).$$

Then for all $t \ge 0$ *the iterates generated by* Byz-DM21 *in Algorithm 1 satisfy*

$$\begin{aligned}\frac{1}{G}\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|M_i^{(t)}\|^2\Big] &= \frac{1}{G}\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|v_i^{(t)}-\nabla f_i(x^{(t)})\|^2\Big]\\ &\le \frac{\widetilde{L}^2}{\eta^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|x^{(t+1)}-x^{(t)}\|^2\right] + \eta T\sigma^2 + \frac{1}{\eta G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\|v_i^{(0)}-\nabla f_i(x^{(0)})\|^2\right],\end{aligned} \tag{25}$$

and

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{E}\Big[\|\widetilde{M}_i^{(t)}\|^2\Big] &= \sum_{t=0}^{T-1}\mathbb{E}\Big[\|\overline{v}^{(t)}-\nabla f(x^{(t)})\|^2\Big]\\ &\le \frac{L^2}{\eta^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|x^{(t+1)}-x^{(t)}\|^2\right] + \frac{\eta T\sigma^2}{G} + \frac{1}{\eta}\mathbb{E}\left[\|\overline{v}^{(0)}-\nabla f_{\mathcal{G}}(x^{(0)})\|^2\right].\end{aligned} \tag{26}$$

Proof. By the update rule of $v_i^{(t)}$, and consider

$$\begin{aligned}&\|v_i^{(t)}-\nabla f_i(x^{(t)})\|^2\\ &= \|(1-\eta)v_i^{(t-1)} + \eta\nabla f_i(x_i^{(t)},\xi_i^{(t)}) - \nabla f_i(x^{(t)})\|^2\\ &= \|(1-\eta)(v_i^{(t-1)}-\nabla f_i(x^{(t)})) + \eta(\nabla f_i(x_i^{(t)},\xi_i^{(t)}) - \nabla f_i(x^{(t)}))\|^2.\end{aligned}$$

Taking the expectation on both sides and using the law of total expectation, we obtain

$$\begin{aligned}&\mathbb{E}\Big[\|v_i^{(t)}-\nabla f_i(x^{(t)})\|^2\Big]\\ &= \mathbb{E}\Big[\mathbb{E}_{\xi_i^{(t)}}\Big[\|(1-\eta)(v_i^{(t-1)}-\nabla f_i(x^{(t)})) + \eta(\nabla f_i(x_i^{(t)},\xi_i^{(t)}) - \nabla f_i(x^{(t)}))\|^2\Big]\Big],\end{aligned}$$

there holds

$$\begin{aligned}&\mathbb{E}\Big[\|v_i^{(t)}-\nabla f_i(x^{(t)})\|^2\Big]\\ &= (1-\eta)^2\mathbb{E}\Big[\|v_i^{(t-1)}-\nabla f_i(x^{(t)})\|^2\Big] + \eta^2\mathbb{E}\Big[\|\nabla f_i(x^{(t)},\xi_i^{(t)}) - \nabla f_i(x^{(t)})\|^2\Big]\\ &\overset{(i)}{\le} (1-\eta)^2(1+a)\mathbb{E}\left[\|v_i^{(t-1)}-\nabla f_i(x^{(t-1)})\|^2\right] + \eta^2\sigma^2\\ &\quad + (1-\eta)^2(1+a^{-1})\mathbb{E}\left[\|\nabla f_i(x^{(t)}) - \nabla f_i(x^{(t-1)})\|^2\right].\end{aligned}$$

where (i) uses Assumption 2.4 and inequality (17). for any $a > 0$, we take $a = \eta(1-\eta)^{-1}$ and use the L-smoothness of $f_i(\cdot)$ to obtain

$$\begin{aligned}
&\mathbb{E}\Big[\|M_i^{(t)}\|^2\Big] \\
&= \mathbb{E}\Big[\|v_i^{(t)} - \nabla f_i(x^{(t)})\|^2\Big] \\
&\leq (1-\eta)\mathbb{E}\left[\|M_i^{(t-1)}\|^2\right] + \frac{L_i^2}{\eta}\mathbb{E}\left[\|x^{(t)} - x^{(t-1)}\|^2\right] + \eta^2\sigma^2.
\end{aligned}$$

Summing up the above inequality over all $i \in \mathcal{G}$ and from $t = 0$ to $t = T-1$ yields

$$\begin{aligned}
&\frac{1}{G}\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|M_i^{(t)}\|^2\Big] \\
&\leq \frac{\widetilde{L}^2}{\eta^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|x^{(t+1)} - x^{(t)}\|^2\right] + \eta T\sigma^2 + \frac{1}{\eta G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\|v_i^{(0)} - \nabla f_i(x^{(0)})\|^2\right].
\end{aligned}$$

Using the same arguments, we obtain

$$\begin{aligned}
&\mathbb{E}\Big[\|\overline{v}^{(t)} - \nabla f_{\mathcal{G}}(x^{(t)})\|^2\Big] \\
&\quad\leq (1-\eta)\mathbb{E}\Big[\|\overline{v}^{(t-1)} - \nabla f_{\mathcal{G}}(x^{(t-1)})\|^2\Big] + \frac{L^2}{\eta}\mathbb{E}\Big[\|x^{(t-1)} - x^{(t)}\|^2\Big] + \frac{\eta^2\sigma^2}{G},
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\Big[\|\overline{v}^{(t)} - \nabla f_{\mathcal{G}}(x^{(t)})\|^2\Big] \\
&\leq \frac{L^2}{\eta^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|x^{(t+1)} - x^{(t)}\|^2\right] + \frac{\eta T\sigma^2}{G} + \frac{1}{\eta}\mathbb{E}\left[\|\overline{v}^{(0)} - \nabla f_{\mathcal{G}}(x^{(0)})\|^2\right].
\end{aligned}$$

□

F.2 Proof of Theorem 3.1

Proof. By Lemma F.1, there holds, for any $\gamma \leq 1/(2L)$,

$$f(x^{(t+1)}) \leq f(x^{(t)}) - \frac{\gamma}{2}\|\nabla f(x^{(t)})\|^2 - \frac{1}{4\gamma}\|x^{(t+1)} - x^{(t)}\|^2 + \frac{\gamma}{2}\|g^{(t)} - \nabla f(x^{(t)})\|^2. \tag{27}$$

Summing the above from $t = 0$ to $t = T-1$ and taking expectation, we define $\delta_t \stackrel{\text{def}}{=} \mathbb{E}\Big[\nabla f(x^{(t)}) - f(x^*)\Big]$, then we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\|\nabla f(x^{(t)})\|^2\Big] \leq \frac{2\delta}{\gamma T} - \frac{1}{2\gamma^2 T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|x^{(t+1)} - x^{(t)}\|^2\right] + \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|g^{(t)} - \nabla f(x^{(t)})\|^2\right].$$

In order to control the error between $g^{(t)}$ and $\nabla f(x^{(t)})$, we decompose it into four terms

$$\begin{aligned}
\|g^{(t)} - \nabla f(x^{(t)})\|^2 &\leq 4\|g^{(t)} - \overline{g}^{(t)}\|^2 + 4\|\overline{g}^{(t)} - \overline{u}^{(t)}\|^2 + 4\|\overline{u}^{(t)} - \overline{v}^{(t)}\|^2 + 4\|\overline{v}^{(t)} - \nabla f(x^{(t)})\|^2 \\
&\leq 4\|g^{(t)} - \overline{g}^{(t)}\|^2 + \frac{4}{G}\sum_{i\in\mathcal{G}}\|g_i^{(t)} - u_i^{(t)}\|^2 \\
&\quad + \frac{4}{G}\sum_{i\in\mathcal{G}}\|u_i^{(t)} - v_i^{(t)}\|^2 + 4\|\overline{v}^{(t)} - \nabla f(x^{(t)})\|^2,
\end{aligned}$$

where $\overline{g} = G^{-1}\sum_{i\in\mathcal{G}} g_i$, $\overline{u} = G^{-1}\sum_{i\in\mathcal{G}} u_i$ and $\overline{v} = G^{-1}\sum_{i\in\mathcal{G}} v_i$.

Next, we apply the technical lemmas from the previous section to derive a bound on the deviation between $g^{(t)}$ and $\nabla f(x^{(t)})$. First, we invoke Lemma F.2 to obtain the following result

$$\begin{aligned}
&\|g^{(t)} - \nabla f(x^{(t)})\|^2 \\
&\leq \frac{32\kappa(G-1)}{G^2}\sum_{i\in\mathcal{G}}\Big(\|C_i^{(t)}\|^2 + \|P_i^{(t)}\|^2 + \|M_i^{(t)}\|^2\Big) + \frac{32\kappa(G-1)}{G}\zeta^2 \\
&\quad + \frac{4}{G}\sum_{i\in\mathcal{G}}\|C_i^{(t)}\|^2 + \frac{4}{G}\sum_{i\in\mathcal{G}}\|P_i^{(t)}\|^2 + 4\|\widetilde{M}_i^{(t)}\|^2 \\
&\leq \frac{4(8\kappa+1)}{G}\sum_{i\in\mathcal{G}}\|C_i^{(t)}\|^2 + \frac{4(8\kappa+1)}{G}\sum_{i\in\mathcal{G}}\|P_i^{(t)}\|^2 + \frac{32\kappa}{G}\sum_{i\in\mathcal{G}}\|M_i^{(t)}\|^2 \\
&\quad + 4\|\widetilde{M}_i^{(t)}\|^2 + 32\kappa\zeta^2,
\end{aligned} \tag{28}$$

where $C_i^{(t)} = g_i^{(t)} - u_i^{(t)}$, $P_i^{(t)} = u_i^{(t)} - v_i^{(t)}$, $\widetilde{M}_i^{(t)} = \overline{v}_i^{(t)} - \nabla f_i(x^{(t)})$. By summing up the above inequality from $t = 0$ to $t = T-1$, we obtain

$$\begin{aligned}
&\sum_{t=0}^{T-1}\|g^{(t)} - \nabla f(x^{(t)})\|^2 \\
&\leq \frac{4(8\kappa+1)}{G}\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\|C_i^{(t)}\|^2 + \frac{4(8\kappa+1)}{G}\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\|P_i^{(t)}\|^2 + \frac{32\kappa}{G}\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\|M_i^{(t)}\|^2 \\
&\quad + 4\sum_{t=0}^{T-1}\|\widetilde{M}_i^{(t)}\|^2 + 32\kappa T\zeta^2.
\end{aligned}$$

Next, by taking the expectation and using Lemma F.3, and define $R^{(t)} \stackrel{\text{def}}{=} \mathbb{E}\Big[\|x^{(t+1)} - x^{(t)}\|^2\Big]$, we have

$$\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\Big[\|g^{(t)} - \nabla f(x^{(t)})\|^2\Big] \\
&\leq \frac{4(8\kappa+1)}{G}\sum_{i\in\mathcal{G}}\Big(\frac{12\eta^4}{\alpha^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|M_i^{(t)}\|^2\right] + \frac{2\eta^4T\sigma^2}{\alpha} + \frac{12\eta^4L_i^2}{\alpha^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|R^{(t)}\|^2\right] + \frac{12\eta^2}{\alpha^2}\mathbb{E}\left[\|P_i^{(t)}\|^2\right]\Big) \\
&\quad + \frac{4(8\kappa+1)}{G}\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|P_i^{(t)}\|^2\Big] + \frac{32\kappa}{G}\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|M_i^{(t)}\|^2\Big] + 4\sum_{t=0}^{T-1}\left[\|\widetilde{M}_i^{(t)}\|^2\right] + 32\kappa T\zeta^2 \\
&\leq \Big(\frac{48\eta^4(8\kappa+1)}{\alpha^2G} + \frac{32\kappa}{G}\Big)\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|M_i^{(t)}\|^2\Big] + \frac{8(8\kappa+1)\eta^4T\sigma^2}{\alpha} + \frac{48\eta^4\widetilde{L}^2(8\kappa+1)}{\alpha^2}\sum_{t=0}^{T-1}\mathbb{E}\Big[\|R^{(t)}\|^2\Big] \\
&\quad + (\frac{\alpha^2+12\eta^2}{\alpha^2})(\frac{4(8\kappa+1)}{G})\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|P_i^{(t)}\|^2\Big] + 4\sum_{t=0}^{T-1}\mathbb{E}\Big[\|\widetilde{M}_i^{(t)}\|^2\Big] + 32\kappa T\zeta^2.
\end{aligned}$$

Then, by using Lemma F.4, we get

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\Big[\|g^{(t)}-\nabla f(x^{(t)})\|^2\Big]\\
&\le \Big(\frac{48\eta^4(8\kappa+1)}{\alpha^2 G}+\frac{32\kappa}{G}\Big)\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\|M_i^{(t)}\|^2\right]+4\sum_{t=0}^{T-1}\mathbb{E}\left[\|\widetilde{M}_i^{(t)}\|^2\right]+32\kappa T\zeta^2\\
&\quad+\frac{48\eta^4\widetilde{L}^2(8\kappa+1)}{\alpha^2}\sum_{t=0}^{T-1}\mathbb{E}\Big[\|R^{(t)}\|^2\Big]+\frac{8(8\kappa+1)\eta^4T\sigma^2}{\alpha}\\
&\quad+(\frac{\alpha^2+12\eta^2}{\alpha^2})(\frac{4(8\kappa+1)}{G})\sum_{i\in\mathcal{G}}\Big(6\sum_{t=0}^{T-1}\mathbb{E}\Big[\|M_i^{(t)}\|^2\Big]+6L_i^2\sum_{t=0}^{T-1}\mathbb{E}\Big[\|R^{(t)}\|^2\Big]+\eta T\sigma^2\Big)\\
&\le \Big(\frac{(48\eta^4+288\eta^2+24\alpha^2)(8\kappa+1)+32\kappa\alpha^2}{\alpha^2 G}\Big)\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|M_i^{(t)}\|^2\Big]+\frac{8(8\kappa+1)\eta^4T\sigma^2}{\alpha}\\
&\quad+4\sum_{t=0}^{T-1}\mathbb{E}\left[\|\widetilde{M}_i^{(t)}\|^2\right]+(\frac{4(8\kappa+1)(\alpha^2+12\eta^2)}{\alpha^2})\eta T\sigma^2+32\kappa T\zeta^2\\
&\quad+\Big(\frac{24\widetilde{L}^2(8\kappa+1)(12\eta^4+\alpha^2+12\eta^2)}{\alpha^2}\Big)\sum_{t=0}^{T-1}\mathbb{E}\Big[\|R^{(t)}\|^2\Big].
\end{aligned}
$$

Furthermore, by using Lemma F.5, we get

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\Big[\|g^{(t)}-\nabla f(x^{(t)})\|^2\Big]\\
&\le \Big(\frac{(48\eta^4+288\eta^2+24\alpha^2)(8\kappa+1)+32\kappa\alpha^2}{\alpha^2}\Big)\Big(\frac{\widetilde{L}^2}{\eta^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|R^{(t)}\|^2\right]+\eta T\sigma^2+\frac{1}{\eta G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\|M_i^{(0)}\|^2\right]\Big)\\
&\quad+\frac{8(8\kappa+1)\eta^4T\sigma^2}{\alpha}+\frac{4L^2}{\eta^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|R^{(t)}\|^2\right]+\frac{4\eta T\sigma^2}{G}+\frac{4}{\eta}\mathbb{E}\left[\|\widetilde{M}_i^{(0)}\|^2\right]+32\kappa T\zeta^2\\
&\quad+(\frac{4(8\kappa+1)(\alpha^2+12\eta^2)}{\alpha^2})\eta T\sigma^2+\Big(\frac{24\widetilde{L}^2(8\kappa+1)(12\eta^4+\alpha^2+12\eta^2)}{\alpha^2}\Big)\sum_{t=0}^{T-1}\mathbb{E}\Big[\|R^{(t)}\|^2\Big]\\
&\le \Bigg(\frac{\widetilde{L}^2(48\eta^4+288\eta^2+24\alpha^2)(8\kappa+1)+32\kappa\widetilde{L}^2\alpha^2+4\alpha^2L^2}{\alpha^2\eta^2}\\
&\quad+\frac{24\widetilde{L}^2(8\kappa+1)(12\eta^4+\alpha^2+12\eta^2)}{\alpha^2}\Bigg)\sum_{t=0}^{T-1}\mathbb{E}\left[\|R^{(t)}\|^2\right]+32\kappa T\zeta^2+\frac{4}{\eta}\mathbb{E}\Big[\|\widetilde{M}^{(0)}\|^2\Big]\\
&\quad+\frac{(48\eta^4+288\eta^2+24\alpha^2)(8\kappa+1)+32\kappa\alpha^2}{\eta\alpha^2G}\mathbb{E}\sum_{i\in\mathcal{G}}\left[\|M_i^{(0)}\|^2\right]\\
&\quad+\Big(\frac{(48\eta^4+336\eta^2+28\alpha^2)(8\kappa+1)}{\alpha^2}+32\kappa+\frac{4}{G}+\frac{8(8\kappa+1)\eta^3}{\alpha}\Big)\eta T\sigma^2.
\end{aligned}
$$

Subtracting $f(x^*)$ from both sides of inequality (27), taking expectation and defining $\delta_t \stackrel{\text{def}}{=} \nabla f(x^{(t)})-f(x^*)$, we

derive

$$\mathbb{E}\left[\|\nabla f(\hat{x}^{(T)})\|^2\right] \leq \frac{2\delta}{\gamma T} - \frac{A}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|x^{(t+1)}-x^{(t)}\|^2\right] + 32\kappa\zeta^2 + \frac{4}{\eta T}\mathbb{E}\Big[\|\widetilde{M}^{(0)}\|^2\Big] + \frac{(48\eta^4+288\eta^2+24\alpha^2)(8\kappa+1)+32\kappa\alpha^2}{\eta\alpha^2 GT}\mathbb{E}\sum_{i\in\mathcal{G}}\left[\|M_i^{(0)}\|^2\right] + \Big(\frac{48(\eta^4+7\eta^2)(8\kappa+1)}{\alpha^2} + 4(64\kappa+7) + \frac{4}{G} + \frac{8(8\kappa+1)\eta^3}{\alpha}\Big)\eta\sigma^2,$$

where $\hat{x}^{(T)}$ is sampled uniformly at random from T iterates and

$$\begin{aligned}
A &= \frac{1}{\gamma^2}\Bigg(\frac{1}{2} - \frac{\gamma^2\widetilde{L}^2(48\eta^4+288\eta^2+24\alpha^2)(8\kappa+1)}{\alpha^2\eta^2} - \frac{4\gamma^2(8\kappa\widetilde{L}^2+L^2)}{\eta^2} \\
&\qquad - \frac{24\gamma^2\widetilde{L}^2(8\kappa+1)(12\eta^4+\alpha^2+12\eta^2)}{\alpha^2}\Bigg) \\
&= \frac{1}{\gamma^2}\Bigg(\frac{1}{2} - \frac{48\gamma^2\widetilde{L}^2(\eta^2+6)(8\kappa+1)}{\alpha^2} - \frac{24\gamma^2\widetilde{L}^2(8\kappa+1)}{\eta^2} - \frac{4\gamma^2(8\kappa\widetilde{L}^2+L^2)}{\eta^2} \\
&\qquad - \frac{24\gamma^2\widetilde{L}^2(8\kappa+1)(12\eta^4+\alpha^2+12\eta^2)}{\alpha^2}\Bigg) \\
&= \frac{1}{\gamma^2}\Bigg(\frac{1}{2} - \frac{48\gamma^2\widetilde{L}^2(\eta^2+6)(8\kappa+1)}{\alpha^2} - \frac{4\gamma^2((56\kappa+6)\widetilde{L}^2+L^2)}{\eta^2} \\
&\qquad - \frac{24\gamma^2\widetilde{L}^2(8\kappa+1)(12\eta^4+\alpha^2+12\eta^2)}{\alpha^2}\Bigg) \\
&= \frac{1}{\gamma^2}\Big(\frac{1}{2} - \frac{24\gamma^2\widetilde{L}^2(8\kappa+1)(12\eta^4+\alpha^2+14\eta^2+12)}{\alpha^2} - \frac{4\gamma^2((56\kappa+6)\widetilde{L}^2+L^2)}{\eta^2}\Big) \\
&\overset{(i)}{\geq} \frac{1}{\gamma^2}\Big(\frac{1}{2} - \frac{936\gamma^2\widetilde{L}^2(8\kappa+1)}{\alpha^2} - \frac{4\gamma^2((56\kappa+6)\widetilde{L}^2+L^2)}{\eta^2}\Big) \\
&\overset{(ii)}{\geq} 0\,,
\end{aligned}$$

where (i) and (ii) are due to $\eta \leq 1$ and the assumption on step-size.

Finally, by using the choice of the momentum parameter, we derive

$$\eta \leq \min\Bigg\{\Big(\frac{8\alpha^2\sqrt{(56\kappa+6)\widetilde{L}^2+L^2}\delta_0}{48(8\kappa+1)\sigma^2T}\Big)^{1/6}, \Big(\frac{8\alpha^2\sqrt{(56\kappa+6)\widetilde{L}^2+L^2}\delta_0}{336(8\kappa+1)\sigma^2T}\Big)^{1/4}, \Big(\frac{8\sqrt{(56\kappa+6)\widetilde{L}^2+L^2}\delta_0}{(64\kappa+7)\sigma^2T}\Big)^{1/2}, \Big(\frac{8\sqrt{(56\kappa+6)\widetilde{L}^2+L^2}\delta_0 G}{4\sigma^2T}\Big)^{1/2}, \Big(\frac{8\alpha\sqrt{(56\kappa+6)\widetilde{L}^2+L^2}\delta_0}{8(8\kappa+1)\sigma^2T}\Big)^{1/5}\Bigg\},$$

ensures that $\frac{48\eta^5(8\kappa+1)\sigma^2}{\alpha^2} \leq \frac{8\sqrt{(56\kappa+6)\widetilde{L}^2+L^2}}{\eta T}$, $\frac{336\eta^3(8\kappa+1)\sigma^2}{\alpha^2} \leq \frac{8\sqrt{(56\kappa+6)\widetilde{L}^2+L^2}}{\eta T}$, $4\eta(64\kappa+7)\sigma^2 \leq \frac{8\sqrt{(56\kappa+6)\widetilde{L}^2+L^2}}{\eta T}$, $\frac{4\eta\sigma^2}{G} \leq \frac{8\sqrt{(56\kappa+6)\widetilde{L}^2+L^2}}{\eta T}$, and $\frac{8\eta^4(8\kappa+1)\sigma^2}{\alpha} \leq \frac{8\sqrt{(56\kappa+6)\widetilde{L}^2+L^2}}{\eta T}$, we denote $\widehat{L} = \sqrt{(\kappa+1)\widetilde{L}^2+L^2}$, then we obtain

$$\mathbb{E}\Big[\|\nabla f(\hat{x}^{(T)})\|^2\Big] \leq \Big(\frac{(48(8\kappa+1))^{1/5}\sigma^{2/5}\widehat{L}\delta_0}{\alpha^{2/5}T}\Big)^{5/6} + \Big(\frac{(336(8\kappa+1))^{1/3}\sigma^{2/3}\widehat{L}\delta_0}{\alpha^{2/3}T}\Big)^{3/4} + 32\kappa\zeta^2 + \frac{\Phi_0}{\gamma T} + \Big(\frac{4(64\kappa+7)\sigma^2\widehat{L}\delta_0}{T}\Big)^{1/2} + \Big(\frac{16\sigma^2\widehat{L}\delta_0}{GT}\Big)^{1/2} + \Big(\frac{(8(8\kappa+1))^{1/4}\sigma^{1/2}\widehat{L}\delta_0}{\alpha^{1/4}T}\Big)^{4/5}.$$

This concludes the proof. □

G Missing Proofs of Byz-VR-DM21 for General Non-Convex Functions

Let us state the following lemma that is used in the analysis of our methods.

G.1 Supporting Lemmas

Lemma G.1 (Robust aggregation error). *Using Lemma F.2 and suppose that Assumption 2.3 holds. Then, for all $t \geq 0$ the iterates generated by* Byz-VR-DM21 *in Algorithm 1 satisfy the following condition:*

$$\begin{aligned}
&\|g^{(t)} - \overline{g}^{(t)}\|^2 \\
&\leq \frac{\kappa}{G}\sum_{i\in\mathcal{G}}\|g_i^{(t)} - \overline{g}^{(t)}\|^2 \\
&\leq \frac{\kappa}{2G^2}\sum_{i,j\in\mathcal{G}}\|g_i^{(t)} - g_j^{(t)}\|^2 \\
&\leq \frac{8\kappa(G-1)}{G^2}\sum_{i\in\mathcal{G}}\left(\|C_i^{(t)}\|^2 + \|P_i^{(t)}\|^2 + \|M_i^{(t)}\|^2\right) + \frac{8\kappa(G-1)}{G}\zeta^2,
\end{aligned}$$

where $\overline{g}^{(t)} = G^{-1}\sum_{i\in\mathcal{G}} g_i$, $C_i^{(t)} = g_i^{(t)} - u_i^{(t)}$, $P_i^{(t)} = u_i^{(t)} - v_i^{(t)}$, $M_i^{(t)} = v_i^{(t)} - \nabla f_i(x^{(t)})$.

Lemma G.2 (Accumulated compression error). *Let Assumption 2.1 and 2.4 be satisfied, and suppose $\mathcal{C}$ is a contractive compressor. For every $i = 1, \ldots, G$, let the sequences $\{v_i^{(t)}\}_{t\geq 0}$, $\{u_i^{(t)}\}_{t\geq 0}$, and $\{g_i^{(t)}\}_{t\geq 0}$ be updated via*

$$\begin{aligned}
g_i^{(t)} &= g_i^{(t-1)} + \mathcal{C}(u_i^{(t)} - g_i^{(t-1)}) \\
u_i^{(t)} &= u_i^{(t-1)} + \eta(v_i^{(t)} - u_i^{(t-1)}) \\
v_i^{(t)} &= (1-\eta)v_i^{(t-1)} + \eta\nabla f_i(x^{(t)}, \xi_i^{(t)}) \\
&\quad + (1-\eta)(\nabla f_i(x^{(t)}, \xi_i^{(t)}) - \nabla f_i(x^{(t-1)}, \xi_i^{(t)})).
\end{aligned}$$

Then for all $t \geq 0$ the iterates generated by Byz-VR-DM21 *in Algorithm 1 satisfy*

$$\begin{aligned}
\sum_{t=0}^{T-1}\mathbb{E}\left[\|C_i^{(t)}\|^2\right] &= \sum_{t=0}^{T-1}\mathbb{E}\left[\|g_i^{(t)} - u_i^{(t)}\|^2\right] \\
&\leq \frac{12\eta^4}{\alpha^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f_i(x^{(t)}) - v_i^{(t)}\|^2\right] + \frac{4\eta^4 T\sigma^2}{\alpha} \\
&\quad + \frac{4\eta^2(\ell_i^2 + \frac{3}{\alpha}L_i^2)}{\alpha}\sum_{t=0}^{T-1}\mathbb{E}\left[\|x^{(t+1)} - x^{(t)}\|^2\right] \\
&\quad + \frac{12\eta^2}{\alpha^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|u_i^{(t)} - v_i^{(t)}\|^2\right].
\end{aligned} \tag{29}$$

Proof. We define

$$\begin{aligned}
\mathcal{S}_i^{(t)} &\overset{\text{def}}{=} \nabla f_i(x^{(t)}, \xi_i^{(t)}) - \nabla f_i(x^{(t)}), \\
\mathcal{S}^{(t)} &\overset{\text{def}}{=} \frac{1}{G}\sum_{i\in\mathcal{G}}\mathcal{S}_i^{(t)}, \\
\mathcal{M}_i^{(t)} &\overset{\text{def}}{=} \nabla f_i(x^{(t)}, \xi_i^{(t)}) - \nabla f_i(x^{(t)}) + \nabla f_i(x^{(t-1)}) - \nabla f_i(x^{(t-1)}, \xi_i^{(t)}), \\
\mathcal{M}^{(t)} &\overset{\text{def}}{=} \frac{1}{G}\sum_{i\in\mathcal{G}}\mathcal{M}_i^{(t)}.
\end{aligned} \tag{30}$$

Then by Assumptions 2.4, we have

$$\begin{aligned}&\mathbb{E}\Big[\mathcal{S}_i^{(t)}\Big] = \mathbb{E}\Big[\mathcal{M}_i^{(t)}\Big] = \mathbb{E}\Big[\mathcal{S}^{(t)}\Big] = \mathbb{E}\Big[\mathcal{M}^{(t)}\Big] = 0,\\ &\mathbb{E}\Big[\|\mathcal{S}_i^{(t)}\|^2\Big] \leq \sigma^2 \quad , \quad \mathbb{E}\Big[\|\mathcal{S}^{(t)}\|^2\Big] \leq \frac{\sigma^2}{G}.\end{aligned} \tag{31}$$

Furthermore, we can derive

$$\begin{aligned}\mathbb{E}\Big[\|\mathcal{M}^{(t)}\|^2\Big] &= \mathbb{E}\Big[\|\frac{1}{G}\sum_{i\in\mathcal{G}}\mathcal{M}_i^{(t)}\|^2\Big]\\ &= \frac{1}{G^2}\mathbb{E}\Big[\|\sum_{i\in\mathcal{G}}\mathcal{M}_i^{(t)}\|^2\Big]\\ &= \frac{1}{G^2}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|\mathcal{M}_i^{(t)}\|^2\Big] + \frac{1}{G^2}\sum_{i,j\in\mathcal{G},i\neq j}\mathbb{E}\Big[\langle\mathcal{M}_i^{(t)},\mathcal{M}_j^{(t)}\rangle\Big]\\ &\overset{(i)}{=} \frac{1}{G^2}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|\mathcal{M}_i^{(t)}\|^2\Big] + \frac{1}{G^2}\sum_{i,j\in\mathcal{G},i\neq j}\Big\langle\mathbb{E}[\mathcal{M}_i^{(t)}],\mathbb{E}[\mathcal{M}_j^{(t)}]\Big\rangle\\ &= \frac{1}{G^2}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|\mathcal{M}_i^{(t)}\|^2\Big]\\ &\leq \frac{1}{G^2}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|\nabla f_i(x^{(t)},\xi_i^{(t)}) - \nabla f_i(x^{(t-1)},\xi_i^{(t)})\|^2\Big]\\ &\leq \frac{1}{G^2}\sum_{i\in\mathcal{G}}\ell_i^2\mathbb{E}\Big[\|x^{(t)}-x^{(t-1)}\|^2\Big] = \frac{\widetilde{\ell}^2}{G}\mathbb{E}\Big[\|x^{(t)}-x^{(t-1)}\|^2\Big].\end{aligned} \tag{32}$$

Step (i) holds due to the conditional independence of $\mathcal{M}_i^{(t)}$ and $\mathcal{M}_j^{(t)}$, while the final inequality follows from the smoothness of the stochastic functions, as stated in Assumption 2.1. Consequently, we obtain the following result:

$$\mathbb{E}\Big[\|\mathcal{M}_i^{(t)}\|^2\Big] \leq \ell_i^2\mathbb{E}\Big[\|x^{(t)}-x^{(t-1)}\|^2\Big], \qquad \mathbb{E}\Big[\|\mathcal{M}^{(t)}\|^2\Big] \leq \frac{\widetilde{\ell}^2}{G}\mathbb{E}\Big[\|x^{(t)}-x^{(t-1)}\|^2\Big], \tag{33}$$

where the first inequality is obtained by using a similar derivation.

Then by the update rules of $g_i^{(t)}$, $u_i^{(t)}$ and $v_i^{(t)}$, we derive

$$
\begin{aligned}
&\mathbb{E}\left[\|g_i^{(t)} - u_i^{(t)}\|^2\right] \\
&= \mathbb{E}\left[\|g_i^{(t-1)} - u_i^{(t)} + \mathcal{C}(u_i^{(t)} - g_i^{(t-1)})\|^2\right] \\
&= \mathbb{E}\left[\mathbb{E}_{\mathcal{C}}\left[\|u_i^{(t)} - g_i^{(t-1)} - \mathcal{C}(u_i^{(t)} - g_i^{(t-1)})\|^2\right]\right] \\
&\overset{(i)}{\leq} (1-\alpha)\mathbb{E}\left[\|u_i^{(t)} - g_i^{(t-1)}\|^2\right] \\
&= (1-\alpha)\mathbb{E}\left[\|u_i^{(t-1)} - g_i^{(t-1)} + \eta(v_i^{(t)} - u_i^{(t-1)})\|^2\right] \\
&= (1-\alpha)\mathbb{E}\left[\|u_i^{(t-1)} - g_i^{(t-1)} + \eta\Big((1-\eta)(v_i^{(t-1)} - \nabla f_i(x^{(t-1)}, \xi^{(t)})) + \nabla f_i(x^{(t)}, \xi^{(t)}) - u_i^{(t-1)}\Big)\|^2\right] \\
&= (1-\alpha)\mathbb{E}\Bigg[\|\eta\Big(v_i^{(t-1)} - u_i^{(t-1)} + \eta(\nabla f_i(x^{(t)}, \xi^{(t)}) - v_i^{(t-1)}) + (1-\eta)(\nabla f_i(x^{(t)}) - \nabla f_i(x^{(t-1)}))\Big) \\
&\quad + \eta(1-\eta)(\nabla f_i(x^{(t)}, \xi_i^{(t)}) - \nabla f_i(x^{(t)}) + \nabla f_i(x^{(t-1)}) - \nabla f_i(x^{(t-1)}, \xi_i^{(t)})) + u_i^{(t-1)} - g_i^{(t-1)}\|^2\Bigg] \\
&= (1-\alpha)\mathbb{E}\Bigg[\|\eta\Big(v_i^{(t-1)} - u_i^{(t-1)} + \eta(\nabla f_i(x^{(t)}, \xi^{(t)}) - \nabla f_i(x^{(t)})) + \eta(\nabla f_i(x^{(t-1)}) - v_i^{(t-1)})\Big) \\
&\quad + u_i^{(t-1)} - g_i^{(t-1)} + \eta(\nabla f_i(x^{(t)}) - \nabla f_i(x^{(t-1)})) \\
&\quad + \eta(1-\eta)(\nabla f_i(x^{(t)}, \xi_i^{(t)}) - \nabla f_i(x^{(t)}) + \nabla f_i(x^{(t-1)}) - \nabla f_i(x^{(t-1)}, \xi_i^{(t)}))\|^2\Bigg] \\
&= (1-\alpha)\mathbb{E}\Bigg[\mathbb{E}_{\xi_i^{(t)}}\Big[\|u_i^{(t-1)} - g_i^{(t-1)} + \eta(v_i^{(t-1)} - u_i^{(t-1)}) + \eta^2\mathcal{S}_i^{(t)} + \eta^2(\nabla f_i(x^{(t-1)}) - v_i^{(t-1)}) \\
&\quad + \eta(\nabla f_i(x^{(t)}) - \nabla f_i(x^{(t-1)})) + \eta(1-\eta)\mathcal{M}_i^{(t)}\|^2\Big]\Bigg] \\
&= (1-\alpha)\mathbb{E}\Bigg[\|u_i^{(t-1)} - g_i^{(t-1)} + \eta(v_i^{(t-1)} - u_i^{(t-1)}) + \eta^2(\nabla f_i(x^{(t-1)}) - v_i^{(t-1)}) \\
&\quad + \eta(\nabla f_i(x^{(t)}) - \nabla f_i(x^{(t-1)}))\|^2\Bigg] + (1-\alpha)\mathbb{E}\Big[\|\eta^2\mathcal{S}_i^{(t)} + \eta(1-\eta)\mathcal{M}_i^{(t)}\|^2\Big] \\
&\leq (1-\alpha)\mathbb{E}\Bigg[\|u_i^{(t-1)} - g_i^{(t-1)} + \eta(v_i^{(t-1)} - u_i^{(t-1)}) + \eta^2(\nabla f_i(x^{(t-1)}) - v_i^{(t-1)}) \\
&\quad + \eta(\nabla f_i(x^{(t)}) - \nabla f_i(x^{(t-1)}))\|^2\Bigg] \\
&\quad + 2(1-\alpha)\eta^4\mathbb{E}\Big[\|\mathcal{S}_i^{(t)}\|^2\Big] + 2(1-\alpha)\eta^2(1-\eta)^2\mathbb{E}\Big[\|\mathcal{M}_i^{(t)}\|^2\Big] \\
&\overset{(ii)}{\leq} (1-\alpha)(1+\rho)\mathbb{E}\left[\|u_i^{(t-1)} - g_i^{(t-1)}\|^2\right] + 2\eta^4\sigma^2 + 2\eta^2\ell_i^2\mathbb{E}\Big[\|x^{(t)} - x^{(t-1)}\|^2\Big] \\
&\quad + (1-\alpha)(1+\rho^{-1})\mathbb{E}\Bigg[\|\eta(v_i^{(t-1)} - u_i^{(t-1)}) + \eta^2(\nabla f_i(x^{(t-1)}) - v_i^{(t-1)}) \\
&\quad + \eta(\nabla f_i(x^{(t)}) - \nabla f_i(x^{(t-1)}))\|^2\Bigg]
\end{aligned}
$$

$$\begin{aligned}\mathbb{E}\left[\|g_i^{(t)}-u_i^{(t)}\|^2\right] &\overset{(iii)}{\leq} (1-\alpha)(1+\rho)\mathbb{E}\left[\|u_i^{(t-1)}-g_i^{(t-1)}\|^2\right]+2\eta^4\sigma^2+2\eta^2\ell_i^2\mathbb{E}\left[\|x^{(t)}-x^{(t-1)}\|^2\right]\\ &\quad+3\eta^2(1-\alpha)(1+\rho^{-1})\mathbb{E}\left[\|v_i^{(t-1)}-u_i^{(t-1)}\|^2\right]\\ &\quad+3\eta^4(1-\alpha)(1+\rho^{-1})\mathbb{E}\left[\|\nabla f_i(x^{(t-1)})-v_i^{(t-1)}\|^2\right]\\ &\quad+3\eta^2(1-\alpha)(1+\rho^{-1})\mathbb{E}\left[\|\nabla f_i(x^{(t)})-\nabla f_i(x^{(t-1)})\|^2\right],\end{aligned}$$

where (i) refers to the contractive property, as defined in Definition 2.7, (ii) hold by inequality (17) for any $\rho>0$, and (iii) leverages the smoothness property of $f_i(\cdot)$. Setting $\rho=\alpha/2$, we obtain

$$(1-\alpha)(1+\frac{\alpha}{2})=1-\frac{\alpha}{2}-\frac{\alpha^2}{2}\leq 1-\frac{\alpha}{2},$$

and

$$(1-\alpha)(1+\frac{2}{\alpha})=\frac{2}{\alpha}-\alpha-1\leq\frac{2}{\alpha}.$$

There holds

$$\begin{aligned}&\mathbb{E}\left[\|g_i^{(t)}-u_i^{(t)}\|^2\right]\\ &=\mathbb{E}\left[\|g_i^{(t-1)}-u_i^{(t)}+\mathcal{C}(u_i^{(t)}-g_i^{(t-1)})\|^2\right]\\ &\leq\left(1-\frac{\alpha}{2}\right)\mathbb{E}\left[\|u_i^{(t-1)}-g_i^{(t-1)}\|^2\right]+\frac{6\eta^4}{\alpha}\mathbb{E}\left[\|\nabla f_i(x^{(t-1)})-v_i^{(t-1)}\|^2\right]+2\eta^4\sigma^2\\ &\quad+\eta^2(2\ell_i^2+\frac{6}{\alpha}L_i^2)\mathbb{E}\left[\|x^{(t)}-x^{(t-1)}\|^2\right]+\frac{6\eta^2}{\alpha}\mathbb{E}\left[\|u_i^{(t-1)}-v_i^{(t-1)}\|^2\right].\end{aligned}$$

Summing up the above inequality from $t=0$ to $t=T-1$ leads to

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{E}\left[\|g_i^{(t)}-u_i^{(t)}\|^2\right]&\leq\frac{12\eta^4}{\alpha^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f_i(x^{(t)})-v_i^{(t)}\|^2\right]+\frac{4\eta^4T\sigma^2}{\alpha}\\ &\quad+\frac{4\eta^2(\ell_i^2+\frac{3}{\alpha}L_i^2)}{\alpha}\sum_{t=0}^{T-1}\mathbb{E}\left[\|x^{(t+1)}-x^{(t)}\|^2\right]\\ &\quad+\frac{12\eta^2}{\alpha^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|u_i^{(t)}-v_i^{(t)}\|^2\right].\end{aligned}$$

□

Lemma G.3 (Accumulated second momentum deviation)**.** *Let Assumption 2.1 and 2.4 be satisfied, and suppose $0<\eta\leq 1$. For every $i=1,\ldots,G$, let the sequences $\{v_i^{(t)}\}_{t\geq 0}$ and $\{u_i^{(t)}\}_{t\geq 0}$ be updated via*

$$\begin{aligned}u_i^{(t)}&=u_i^{(t-1)}+\eta(v_i^{(t)}-u_i^{(t-1)}),\\ v_i^{(t)}&=(1-\eta)v_i^{(t-1)}+\eta\nabla f_i(x^{(t)},\xi_i^{(t)})\\ &\quad+(1-\eta)(\nabla f_i(x^{(t)},\xi_i^{(t)})-\nabla f_i(x^{(t-1)},\xi_i^{(t)})).\end{aligned}$$

Then for all $t\geq 0$ the iterates generated by Byz-VR-DM21 *in Algorithm 1 satisfy*

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{E}\left[\|P_i^{(t)}\|^2\right]&=\sum_{t=0}^{T-1}\mathbb{E}\left[\|u_i^{(t)}-v_i^{(t)}\|^2\right]\\ &\leq 6\sum_{t=0}^{T-1}\mathbb{E}\left[\|M_i^{(t)}\|^2\right]+\frac{2(\ell_i^2+\frac{2}{\eta}L_i^2)}{\eta}\sum_{t=0}^{T-1}\mathbb{E}\left[\|x^{(t+1)}-x^{(t)}\|^2\right]+2\eta T\sigma^2.\end{aligned}\tag{34}$$

Proof. By the update rule of $u_i^{(t)}$ and $v_i^{(t)}$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\|u_i^{(t)} - v_i^{(t)}\|^2\right] \\
&= \mathbb{E}\left[\|u_i^{(t-1)} - v_i^{(t)} + \eta(v_i^{(t)} - u_i^{(t-1)})\|^2\right] \\
&= (1-\eta)^2\mathbb{E}\left[\|u_i^{(t-1)} - v_i^{(t)}\|^2\right] \\
&= (1-\eta)^2\mathbb{E}\Big[\|(1-\eta)v_i^{(t-1)} + \eta\nabla f_i(x^{(t)}, \xi_i^{(t)}) - u_i^{(t-1)} + (1-\eta)(\nabla f_i(x^{(t)}, \xi_i^{(t)}) - \nabla f_i(x^{(t-1)}, \xi_i^{(t)}))\|^2\Big] \\
&= (1-\eta)^2\mathbb{E}\Big[\|(v_i^{(t-1)} - u_i^{(t-1)}) + \eta(\nabla f_i(x^{(t)}, \xi_i^{(t)}) - v_i^{(t-1)}) + (1-\eta)(\nabla f_i(x^{(t)}) - \nabla f_i(x^{(t-1)})) \\
&\quad + (1-\eta)(\nabla f_i(x^{(t)}, \xi_i^{(t)}) - \nabla f_i(x^{(t)}) + \nabla f_i(x^{(t-1)}) - \nabla f_i(x^{(t-1)}, \xi_i^{(t)}))\|^2\Big] \\
&= (1-\eta)^2\mathbb{E}\Big[\|(v_i^{(t-1)} - u_i^{(t-1)}) + \eta(\nabla f_i(x^{(t)}, \xi_i^{(t)}) - \nabla f_i(x^{(t)})) + \eta(\nabla f_i(x^{(t-1)}) - v_i^{(t-1)}) \\
&\quad + \nabla f_i(x^{(t)}) - \nabla f_i(x^{(t-1)}) + (1-\eta)\mathcal{M}_i^{(t)}\|^2\Big] \\
&= (1-\eta)^2\mathbb{E}\Big[\mathbb{E}_{\xi_i^{(t)}}\Big[\|(v_i^{(t-1)} - u_i^{(t-1)}) + \eta(\nabla f_i(x^{(t)}, \xi_i^{(t)}) - \nabla f_i(x^{(t)})) + \eta(\nabla f_i(x^{(t-1)}) - v_i^{(t-1)}) \\
&\quad + \nabla f_i(x^{(t)}) - \nabla f_i(x^{(t-1)}) + (1-\eta)\mathcal{M}_i^{(t)}\|^2\Big]\Big] \\
&= (1-\eta)^2\mathbb{E}\Big[\|(v_i^{(t-1)} - u_i^{(t-1)}) + \eta(\nabla f_i(x^{(t-1)}) - v_i^{(t-1)}) + (\nabla f_i(x^{(t)}) - \nabla f_i(x^{(t-1)}))\|^2\Big] \\
&\quad + (1-\eta)^2\mathbb{E}\Big[\|\eta\mathcal{S}_i^t + (1-\eta)\mathcal{M}_i^t\|^2\Big] \\
&\overset{(i)}{\leq} (1-\eta)^2(1+\rho)\mathbb{E}\Big[\|u_i^{(t-1)} - v_i^{(t-1)}\|^2\Big] + (1-\eta)^2(1+\rho^{-1})\mathbb{E}\Bigg[\|\eta(v_i^{(t-1)} - \nabla f_i(x^{(t-1)})) \\
&\quad + \nabla f_i(x^{(t)}) - \nabla f_i(x^{(t-1)})\|^2\Bigg] + 2(1-\eta)^2\eta^2\mathbb{E}\Big[\|\mathcal{S}_i^t\|^2\Big] + 2(1-\eta)^4\mathbb{E}\Big[\|\mathcal{M}_i^t\|^2\Big] \\
&\leq (1-\eta)^2(1+\rho)\mathbb{E}\Big[\|u_i^{(t-1)} - v_i^{(t-1)}\|^2\Big] + 2\eta^2(1+\rho^{-1})\mathbb{E}\Big[\|v_i^{(t-1)} - \nabla f_i(x^{(t-1)})\|^2\Big] + 2\eta^2\sigma^2 \\
&\quad + 2(1-\eta)^2(1+\rho^{-1})\mathbb{E}\Big[\|\nabla f_i(x^{(t)}) - \nabla f_i(x^{(t-1)})\|^2\Big] + 2\ell_i^2\mathbb{E}\Big[\|x^{(t)} - x^{(t-1)}\|^2\Big],
\end{aligned}
$$

where (i) uses inequality (32), smoothness of $f_i(\cdot)$ and inequality (17) for any $\rho > 0$. Setting $\rho = \eta/2$, and because of $\eta \in (0, 1]$, we obtain

$$
(1-\eta)^2(1+\frac{\eta}{2}) = 1 - \frac{3\eta}{2} + \frac{\eta^3}{2} \leq 1 - \eta, \quad (1-\eta)^2(1+\frac{2}{\eta}) \leq 1 + \frac{2}{\eta} - \eta - 2 \leq \frac{2}{\eta}
$$

and

$$
\eta^2(1+\frac{2}{\eta}) = \eta^2 + 2\eta \leq 3\eta.
$$

There holds

$$
\begin{aligned}
&\mathbb{E}\left[\|u_i^{(t)} - v_i^{(t)}\|^2\right] \\
&\leq (1-\eta)^2(1+\frac{\eta}{2})\mathbb{E}\Big[\|u_i^{(t-1)} - v_i^{(t-1)}\|^2\Big] + 2\eta^2(1+\frac{2}{\eta})\mathbb{E}\Big[\|v_i^{(t-1)} - \nabla f_i(x^{(t-1)})\|^2\Big] \\
&\quad + 2(1-\eta)^2(1+\frac{2}{\eta})\mathbb{E}\Big[\|\nabla f_i(x^{(t)}) - \nabla f_i(x^{(t-1)})\|^2\Big] + 2\eta^2\sigma^2 + 2\ell_i^2\mathbb{E}\Big[\|x^{(t)} - x^{(t-1)}\|^2\Big] \\
&\leq (1-\eta)\mathbb{E}\Big[\|u_i^{(t-1)} - v_i^{(t-1)}\|^2\Big] + 6\eta\mathbb{E}\Big[\|v_i^{(t-1)} - \nabla f_i(x^{(t-1)})\|^2\Big] \\
&\quad + 2(\ell_i^2 + \frac{2}{\eta}L_i^2)\mathbb{E}\Big[\|x^{(t)} - x^{(t-1)}\|^2\Big] + 2\eta^2\sigma^2.
\end{aligned}
$$

Summing up the above inequality from $t=0$ to $t=T-1$ yields

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{E}\left[\|P_i^{(t)}\|^2\right] &\le 6\sum_{t=0}^{T-1}\mathbb{E}\Big[\|v_i^{(t)}-\nabla f_i(x^{(t)}\|^2\Big]+\frac{2(\ell_i^2+\frac{2}{\eta}L_i^2)}{\eta}\sum_{t=0}^{T-1}\mathbb{E}\Big[\|x^{(t+1)}-x^{(t)}\|^2\Big]\\ &\quad+\frac{1}{\eta}\sum_{t=0}^{T-1}\mathbb{E}\left[\|P_i^{(0)}\|^2\right]+2\eta T\sigma^2\\ &\le 6\sum_{t=0}^{T-1}\mathbb{E}\Big[\|M_i^{(t)}\|^2\Big]+\frac{2(\ell_i^2+\frac{2}{\eta})L_i^2}{\eta}\sum_{t=0}^{T-1}\mathbb{E}\Big[\|x^{(t+1)}-x^{(t)}\|^2\Big]+2\eta T\sigma^2,\end{aligned}$$

where $P_i^{(0)}=u_i^{(0)}-v_i^{(0)}$, because of $u_i^{(0)}=v_i^{(0)}$, $P_i^{(0)}=0$. □

Lemma G.4 (Accumulated momentum deviation)**.** *Let Assumption 2.1 and 2.4 be satisfied, and suppose $0<\eta\le 1$. For every $i=1,\dots,G$, let the sequences $\{v_i^{(t)}\}_{t\ge 0}$ be updated via*

$$v_i^{(t)}=(1-\eta)v_i^{(t-1)}+\eta\nabla f_i(x^{(t)},\xi_i^{(t)})+(1-\eta)(\nabla f_i(x^{(t)},\xi_i^{(t)})-\nabla f_i(x^{(t-1)},\xi_i^{(t)})).$$

Then for all $t\ge 0$ the iterates generated by Byz-VR-DM21 *in Algorithm 1 satisfy*

$$\begin{aligned}\frac{1}{G}\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|M_i^{(t)}\|^2\Big]&=\frac{1}{G}\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|v_i^{(t)}-\nabla f_i(x^{(t)})\|^2\Big]\\ &\le\frac{2\widetilde{\ell}^2}{\eta}\sum_{t=0}^{T-1}\mathbb{E}\left[\|x^{(t+1)}-x^{(t)}\|^2\right]+2\eta T\sigma^2+\frac{1}{\eta G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\|v_i^{(0)}-\nabla f_i(x^{(0)})\|^2\right],\end{aligned}\tag{35}$$

and

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{E}\Big[\|\widetilde{M_i^{(t)}}\|^2\Big]&=\sum_{t=0}^{T-1}\mathbb{E}\Big[\|\overline{v}^{(t)}-\nabla f_{\mathcal{G}}(x^{(t)})\|^2\Big]\\ &\le\frac{2\widetilde{\ell}^2}{\eta}\sum_{t=0}^{T-1}\mathbb{E}\left[\|x^{(t+1)}-x^{(t)}\|^2\right]+\frac{2\eta T\sigma^2}{G}+\frac{1}{\eta}\mathbb{E}\left[\|\overline{v}^{(0)}-\nabla f_{\mathcal{G}}(x^{(0)})\|^2\right].\end{aligned}\tag{36}$$

Proof. By the update rule of $v_i^{(t)}$, and consider

$$\begin{aligned}&\|v_i^{(t)}-\nabla f_i(x^{(t)})\|^2\\ &=\|(1-\eta)(v_i^{(t-1)}-\nabla f_i(x^{(t)}))+\eta(\nabla f_i(x_i^{(t)},\xi_i^{(t)})-\nabla f_i(x^{(t)}))\\ &\quad+(1-\eta)(\nabla f_i(x^{(t)},\xi_i^{(t)})-\nabla f_i(x^{(t-1)},\xi_i^{(t)}))\|^2\\ &=\|(1-\eta)(v_i^{(t-1)}-\nabla f_i(x^{(t)}))+\eta(\nabla f_i(x_i^{(t-1)},\xi_i^{(t)})-\nabla f_i(x^{(t)}))\\ &\quad+\nabla f_i(x^{(t)},\xi_i^{(t)})-\nabla f_i(x^{(t-1)},\xi_i^{(t)})\|^2\\ &=\|(1-\eta)(v_i^{(t-1)}-\nabla f_i(x^{(t-1)}))+\eta(\nabla f_i(x^{(t)},\xi^{(t)})-\nabla f_i(x^{(t)}))\\ &\quad+(1-\eta)\Big(\nabla f_i(x^{(t-1)})-\nabla f_i(x^{(t-1)},\xi^{(t)})+\nabla f_i(x^{(t)},\xi^{(t)})-\nabla f_i(x^{(t)})\Big)\|^2\\ &=\|(1-\eta)(v_i^{(t-1)}-\nabla f(x^{(t-1)})+\eta\mathcal{S}_i^{(t)}+(1-\eta)\mathcal{M}_i^{(t)}\|^2.\end{aligned}$$

Taking the expectation on both sides and using the law of total expectation, we obtain

$$\begin{aligned}&\mathbb{E}\Big[\|v_i^{(t)}-\nabla f_i(x^{(t)})\|^2\Big]\\ &=\mathbb{E}\Big[\mathbb{E}_{\xi_i^{(t)}}\Big[\|(1-\eta)(v_i^{(t-1)}-\nabla f(x^{(t-1)})+\eta\mathcal{S}_i^{(t)}+(1-\eta)\mathcal{M}_i^{(t)}\|^2\Big]\Big].\end{aligned}$$

Then, there holds

$$\begin{aligned}
&\mathbb{E}\Big[\|v_i^{(t)} - \nabla f_i(x^{(t)})\|^2\Big] \\
&= (1-\eta)^2\mathbb{E}\Big[\|v_i^{(t-1)} - \nabla f_i(x^{(t-1)})\|^2\Big] + \mathbb{E}\Big[\|\eta\mathcal{S}_i^{(t)} + (1-\eta)\mathcal{M}_i^{(t)}\|^2\Big] \\
&\le (1-\eta)\mathbb{E}\left[\|v_i^{(t-1)} - \nabla f_i(x^{(t-1)})\|^2\right] + 2\eta^2\mathbb{E}\left[\|\mathcal{S}_i^{(t)}\|^2\right] + 2\mathbb{E}\left[\|\mathcal{M}_i^{(t)}\|^2\right] \\
&\le (1-\eta)\mathbb{E}\left[\|v_i^{(t-1)} - \nabla f_i(x^{(t-1)})\|^2\right] + 2\eta^2\sigma^2 + 2\ell_i^2\mathbb{E}\left[\|x^{(t)} - x^{(t-1)}\|^2\right].
\end{aligned}$$

Summing up the above inequality over all $i \in \mathcal{G}$ and from $t = 0$ to $t = T-1$ yields

$$\begin{aligned}
&\frac{1}{G}\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|M_i^{(t)}\|^2\Big] \\
&\le \frac{2\widetilde{\ell}^2}{\eta}\sum_{t=0}^{T-1}\mathbb{E}\left[\|x^{(t)} - x^{(t-1)}\|^2\right] + 2\eta T\sigma^2 + \frac{1}{\eta G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\|v_i^{(0)} - \nabla f_i(x^{(0)})\|^2\right].
\end{aligned}$$

Using the same arguments, we obtain

$$\begin{aligned}
&\mathbb{E}\Big[\|\overline{v}^{(t)} - \nabla f_{\mathcal{G}}(x^{(t)})\|^2\Big] \\
&\quad\le (1-\eta)\mathbb{E}\left[\|\overline{v}^{(t-1)} - \nabla f_{\mathcal{G}}(x^{(t-1)})\|^2\right] + \widetilde{\ell}^2\mathbb{E}\left[\|x^{(t)} - x^{(t-1)}\|^2\right] + \frac{2\eta^2\sigma^2}{G}
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\Big[\|\overline{v}^{(t)} - \nabla f_{\mathcal{G}}(x^{(t)})\|^2\Big] \\
&\le \frac{2\widetilde{\ell}^2}{\eta}\sum_{t=0}^{T-1}\mathbb{E}\left[\|x^{(t)} - x^{(t-1)}\|^2\right] + \frac{2\eta T\sigma^2}{G} + \frac{1}{\eta}\mathbb{E}\left[\|\overline{v}^{(0)} - \nabla f_{\mathcal{G}}(x^{(0)})\|^2\right].
\end{aligned}$$

□

G.2 Proof of Theorem 4.1

Proof. By Lemma F.1, there holds, for any $\gamma \le 1/(2L)$,

$$f(x^{(t+1)}) \le f(x^{(t)}) - \frac{\gamma}{2}\|\nabla f(x^{(t)})\|^2 - \frac{1}{4\gamma}\|x^{(t+1)} - x^{(t)}\|^2 + \frac{\gamma}{2}\|g^{(t)} - \nabla f(x^{(t)})\|^2. \tag{37}$$

Summing the above from $t = 0$ to $t = T-1$ and taking expectation, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\|\nabla f(x^{(t)})\|^2\Big] \le \frac{2\delta}{\gamma T} - \frac{1}{2\gamma^2 T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|x^{(t+1)} - x^{(t)}\|^2\right] + \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|g^{(t)} - \nabla f(x^{(t)})\|^2\right],$$

where $\delta_t = \mathbb{E}\Big[\nabla f(x^{(t)}) - f(x^*)\Big]$. We note that

$$\begin{aligned}
\|g^{(t)} - \nabla f(x^{(t)})\|^2 &\le 4\|g^{(t)} - \overline{g}^{(t)}\|^2 + 4\|\overline{g}^{(t)} - \overline{u}^{(t)}\|^2 + 4\|\overline{u}^{(t)} - \overline{v}^{(t)}\|^2 + 4\|\overline{v}^{(t)} - \nabla f(x^{(t)})\|^2 \\
&\le 4\|g^{(t)} - \overline{g}^{(t)}\|^2 + \frac{4}{G}\sum_{i\in\mathcal{G}}\|g_i^{(t)} - u_i^{(t)}\|^2 \\
&\quad + \frac{4}{G}\sum_{i\in\mathcal{G}}\|u_i^{(t)} - v_i^{(t)}\|^2 + 4\|\overline{v}^{(t)} - \nabla f(x^{(t)})\|^2,
\end{aligned}$$

where $\overline{g} = G^{-1}\sum_{i\in\mathcal{G}} g_i$, $\overline{u} = G^{-1}\sum_{i\in\mathcal{G}} u_i$ and $\overline{v} = G^{-1}\sum_{i\in\mathcal{G}} v_i$.

Next, we apply the technical lemmas from the previous section to derive a bound on the deviation between $g^{(t)}$ and $\nabla f(x^{(t)})$. First, we invoke Lemma F.2 to obtain the following result

$$\begin{aligned}
&\|g^{(t)} - \nabla f(x^{(t)})\|^2 \\
&\leq \frac{32\kappa(G-1)}{G^2}\sum_{i\in\mathcal{G}}\left(\|C_i^{(t)}\|^2 + \|P_i^{(t)}\|^2 + \|M_i^{(t)}\|^2\right) + \frac{32\kappa(G-1)}{G}\zeta^2 \\
&\quad + \frac{4}{G}\sum_{i\in\mathcal{G}}\|C_i^{(t)}\|^2 + \frac{4}{G}\sum_{i\in\mathcal{G}}\|P_i^{(t)}\|^2 + 4\|\widetilde{M}_i^{(t)}\|^2 \\
&\leq \frac{4(8\kappa+1)}{G}\sum_{i\in\mathcal{G}}\|C_i^{(t)}\|^2 + \frac{4(8\kappa+1)}{G}\sum_{i\in\mathcal{G}}\|P_i^{(t)}\|^2 + \frac{32\kappa}{G}\sum_{i\in\mathcal{G}}\|M_i^{(t)}\|^2 \\
&\quad + 4\|\widetilde{M}_i^{(t)}\|^2 + 32\kappa\zeta^2,
\end{aligned}$$

where $C_i^{(t)} = g_i^{(t)} - u_i^{(t)}$, $P_i^{(t)} = u_i^{(t)} - v_i^{(t)}$, $\widetilde{M}_i^{(t)} = \overline{v}_i^{(t)} - \nabla f_i(x^{(t)})$. By summing up the above inequality from $t=0$ to $t=T-1$, we obtain

$$\begin{aligned}
&\sum_{t=0}^{T-1}\|g^{(t)} - \nabla f(x^{(t)})\|^2 \\
&\leq \frac{4(8\kappa+1)}{G}\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\|C_i^{(t)}\|^2 + \frac{4(8\kappa+1)}{G}\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\|P_i^{(t)}\|^2 + \frac{32\kappa}{G}\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\|M_i^{(t)}\|^2 \\
&\quad + 4\sum_{t=0}^{T-1}\|\widetilde{M}_i^{(t)}\|^2 + 32\kappa T\zeta^2.
\end{aligned}$$

Next, by taking expectation and using G.2 and define $R^{(t)} \overset{\text{def}}{=} \mathbb{E}\Big[\|x^{(t+1)} - x^{(t)}\|^2\Big]$, we have

$$\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\Big[\|g^{(t)} - \nabla f(x^{(t)})\|^2\Big] \\
&\leq \frac{4(8\kappa+1)}{G}\sum_{i\in\mathcal{G}}\Big(\frac{12\eta^4}{\alpha^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|M_i^{(t)}\|^2\right] + \frac{4\eta^2(\ell_i^2 + \frac{3}{\alpha}L_i^2)}{\alpha}\sum_{t=0}^{T-1}\mathbb{E}\left[\|R^{(t)}\|^2\right] \\
&\quad + \frac{4\eta^4 T\sigma^2}{\alpha} + \frac{12\eta^2}{\alpha^2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|P_i^{(t)}\|^2\right]\Big) + \frac{4(8\kappa+1)}{G}\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|P_i^{(t)}\|^2\Big] \\
&\quad + \frac{32\kappa}{G}\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|M_i^{(t)}\|^2\Big] + 4\sum_{t=0}^{T-1}\left[\|\widetilde{M}_i^{(t)}\|^2\right] + 32\kappa T\zeta^2 \\
&\leq (\frac{48\eta^4(8\kappa+1)}{\alpha^2 G} + \frac{32\kappa}{G})\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|M_i^{(t)}\|^2\Big] + \frac{16(8\kappa+1)\eta^4 T\sigma^2}{\alpha} \\
&\quad + \frac{16\eta^2(8\kappa+1)(\widetilde{\ell}^2 + \frac{3}{\alpha}\widetilde{L}^2)}{\alpha}\sum_{t=0}^{T-1}\mathbb{E}\Big[\|R^{(t)}\|^2\Big] + (\frac{\alpha^2+12\eta^2}{\alpha^2})(\frac{4(8\kappa+1)}{G})\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|P_i^{(t)}\|^2\Big] \\
&\quad + 4\sum_{t=0}^{T-1}\mathbb{E}\Big[\|\widetilde{M}_i^{(t)}\|^2\Big] + 32\kappa T\zeta^2.
\end{aligned}$$

Then, by using G.3, we get

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\Big[\|g^{(t)}-\nabla f(x^{(t)})\|^2\Big]\\
&\le \Big(\frac{48\eta^4(8\kappa+1)}{\alpha^2 G}+\frac{32\kappa}{G}\Big)\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\|M_i^{(t)}\|^2\right]+4\sum_{t=0}^{T-1}\mathbb{E}\left[\|\widetilde{M}_i^{(t)}\|^2\right]+32\kappa T\zeta^2\\
&\quad+\frac{16\eta^2(8\kappa+1)(\widetilde{\ell}^2+\frac{3}{\alpha}\widetilde{L}^2)}{\alpha}\mathbb{E}\Big[\|R^{(t)}\|^2\Big]+\frac{16(8\kappa+1)\eta^4T\sigma^2}{\alpha}\\
&\quad+(\frac{\alpha^2+12\eta^2}{\alpha^2})(\frac{4(8\kappa+1)}{G})\sum_{i\in\mathcal{G}}\Big(6\sum_{t=0}^{T-1}\mathbb{E}\Big[\|M_i^{(t)}\|^2\Big]+\frac{2(\ell_i^2+\frac{2}{\eta}L_i^2)}{\eta}\sum_{t=0}^{T-1}\mathbb{E}\Big[\|R^{(t)}\|^2\Big]+2\eta T\sigma^2\Big)\\
&\le \Big(\frac{48\eta^4(8\kappa+1)+24(8\kappa+1)(\alpha^2+12\eta^2)+32\kappa\alpha^2}{\alpha^2 G}\Big)\sum_{t=0}^{T-1}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|M_i^{(t)}\|^2\Big]+\frac{16(8\kappa+1)\eta^4T\sigma^2}{\alpha}\\
&\quad+4\sum_{t=0}^{T-1}\mathbb{E}\left[\|\widetilde{M}_i^{(t)}\|^2\right]+(\frac{8(\alpha^2+12\eta^2)(8\kappa+1)}{\alpha^2})\eta T\sigma^2+32\kappa T\zeta^2\\
&\quad+\Big(\frac{16\eta^2(8\kappa+1)(\widetilde{\ell}^2+\frac{3}{\alpha}\widetilde{L}^2)}{\alpha}+\frac{8(8\kappa+1)(\alpha^2+12\eta^2)(\widetilde{\ell}^2+\frac{2}{\eta}\widetilde{L}^2)}{\alpha^2\eta}\Big)\sum_{t=0}^{T-1}\mathbb{E}\Big[\|R^{(t)}\|^2\Big].
\end{aligned}
$$

Furthermore, by using G.4, we get

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\Big[\|g^{(t)}-\nabla f(x^{(t)})\|^2\Big]\\
&\le \Big(\frac{48\eta^4(8\kappa+1)+24(8\kappa+1)(\alpha^2+12\eta^2)+32\kappa\alpha^2}{\alpha^2}\Big)\Big(\frac{2\widetilde{\ell}^2}{\eta}\sum_{t=0}^{T-1}\mathbb{E}\left[\|R^{(t)}\|^2\right]+2\eta T\sigma^2\\
&\quad+\frac{1}{\eta G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\|M_i^{(0)}\|^2\right]\Big)+\frac{16(8\kappa+1)\eta^4T\sigma^2}{\alpha}+\frac{8\widetilde{\ell}^2}{\eta}\sum_{t=0}^{T-1}\mathbb{E}\left[\|R^{(t)}\|^2\right]\\
&\quad+\frac{8\eta T\sigma^2}{G}+\frac{4}{\eta}\mathbb{E}\left[\|\widetilde{M}^{(0)}\|^2\right]+(\frac{8(\alpha^2+12\eta^2)(8\kappa+1)}{\alpha^2})\eta T\sigma^2+32\kappa T\zeta^2\\
&\quad+\Big(\frac{16\eta^2(8\kappa+1)(\widetilde{\ell}^2+\frac{3}{\alpha}\widetilde{L}^2)}{\alpha}+\frac{8(8\kappa+1)(\alpha^2+12\eta^2)(\widetilde{\ell}^2+\frac{2}{\eta}\widetilde{L}^2)}{\alpha^2\eta}\Big)\sum_{t=0}^{T-1}\mathbb{E}\Big[\|R^{(t)}\|^2\Big]\\
&\le \Big(\frac{\widetilde{\ell}^2(96\eta^4(8\kappa+1)+48(8\kappa+1)(\alpha^2+12\eta^2)+64\kappa\alpha^2)}{\alpha^2\eta}+\frac{8(8\kappa+1)(\alpha^2+12\eta^2)(\widetilde{\ell}^2+\frac{2}{\eta}\widetilde{L}^2)}{\alpha^2\eta}\\
&\quad+\frac{8\widetilde{\ell}^2}{\eta}+\frac{16\eta^2(8\kappa+1)(\widetilde{\ell}^2+\frac{3}{\alpha}\widetilde{L}^2)}{\alpha}\Big)\sum_{t=0}^{T-1}\mathbb{E}\left[\|R^{(t)}\|^2\right]+32\kappa T\zeta^2+\frac{4}{\eta}\mathbb{E}\Big[\|\widetilde{M}^{(0)}\|^2\Big]\\
&\quad+\frac{48\eta^4(8\kappa+1)+24(8\kappa+1)(\alpha^2+12\eta^2)+32\kappa\alpha^2}{\alpha^2\eta G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\|M_i^{(0)}\|^2\right]\\
&\quad+\Big(\frac{96\eta^4(8\kappa+1)+56(8\kappa+1)(\alpha^2+12\eta^2)}{\alpha^2}+32\kappa+\frac{8}{G}+\frac{16\eta^3(8\kappa+1)}{\alpha}\Big)\eta T\sigma^2.
\end{aligned}
$$

Subtracting $f(x^*)$ from both sides of inequality (37), taking expectation and defining $\delta_t \stackrel{\text{def}}{=} \mathbb{E}\Big[\nabla f(x^{(t)})-f(x^*)\Big]$,

we derive

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\hat{x}^{(T)})\|^2\right] \\
&\leq \frac{2\delta}{\gamma T} - \frac{A}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|x^{(t+1)}-x^{(t)}\|^2\right] + 32\kappa\zeta^2 + \frac{4}{\eta T}\mathbb{E}\Big[\|\widetilde{M}^{(0)}\|^2\Big] \\
&\quad + \frac{48\eta^4(8\kappa+1)+24(8\kappa+1)(\alpha^2+12\eta^2)+32\kappa\alpha^2}{\alpha^2 T\eta G}\sum_{i\in\mathcal{G}}\mathbb{E}\left[\|M_i^{(0)}\|^2\right] \\
&\quad + \Big(\frac{96(8\kappa+1)(\eta^4+7\eta^2)}{\alpha^2} + 8(60\kappa+7) + \frac{8}{G} + \frac{16\eta^3(8\kappa+1)}{\alpha}\Big)\eta\sigma^2,
\end{aligned}$$

where $\hat{x}^{(T)}$ is sampled uniformly at random from T iterates and

$$\begin{aligned}
A &= \frac{1}{\gamma^2}\Big(\frac{1}{2} - \frac{16\gamma^2\widetilde{\ell}^2(6\eta^4(8\kappa+1)+3(8\kappa+1)(\alpha^2+12\eta^2)+4\kappa\alpha^2)}{\alpha^2\eta} - \frac{8\gamma^2\widetilde{\ell}^2}{\eta} \\
&\qquad - \frac{8\gamma^2(8\kappa+1)(\alpha^2+12\eta^2)(\widetilde{\ell}^2+\frac{2}{\eta}\widetilde{L}^2)}{\alpha^2\eta} - \frac{16\gamma^2\eta^2(8\kappa+1)(\widetilde{\ell}^2+\frac{3}{\alpha}\widetilde{L}^2)}{\alpha}\Big) \\
&= \frac{1}{\gamma^2}\Big(\frac{1}{2} - \frac{16\gamma^2\widetilde{\ell}^2(6\eta^4(8\kappa+1)+3(8\kappa+1)(\alpha^2+12\eta^2))}{\alpha^2\eta} - \frac{64\gamma^2\widetilde{\ell}^2\kappa}{\eta} - \frac{8\gamma^2\widetilde{\ell}^2}{\eta} \\
&\qquad - \frac{8\gamma^2\widetilde{\ell}^2(8\kappa+1)(\alpha^2+12\eta^2)}{\alpha^2\eta} - \frac{16\gamma^2\widetilde{L}^2(8\kappa+1)(\alpha^2+12\eta^2)}{\alpha^2\eta^2} \\
&\qquad - \frac{16\gamma^2\widetilde{\ell}^2\eta^2(8\kappa+1)}{\alpha} - \frac{48\gamma^2\widetilde{L}^2\eta^2(8\kappa+1)}{\alpha^2}\Big) \\
&= \frac{1}{\gamma^2}\Big(\frac{1}{2} - \frac{96\gamma^2\widetilde{\ell}^2\eta^3(8\kappa+1)}{\alpha^2} - \frac{48\gamma^2\widetilde{\ell}^2(8\kappa+1)}{\eta} - \frac{576\eta\gamma^2\widetilde{\ell}^2(8\kappa+1)}{\alpha^2} - \frac{64\gamma^2\widetilde{\ell}^2\kappa}{\eta} - \frac{8\gamma^2\widetilde{\ell}^2}{\eta} \\
&\qquad - \frac{8\gamma^2\widetilde{\ell}^2(8\kappa+1)}{\eta} - \frac{96\eta\gamma^2\widetilde{\ell}^2(8\kappa+1)}{\alpha^2} - \frac{16\gamma^2\widetilde{L}^2(8\kappa+1)}{\eta^2} - \frac{192\gamma^2\widetilde{L}^2(8\kappa+1)}{\alpha^2} \\
&\qquad - \frac{16\gamma^2\widetilde{\ell}^2\eta^2(8\kappa+1)}{\alpha} - \frac{48\gamma^2\widetilde{L}^2\eta^2(8\kappa+1)}{\alpha^2}\Big) \\
&= \frac{1}{\gamma^2}\Big(\frac{1}{2} - \frac{8\gamma^2(8\kappa+1)(12\eta^3\widetilde{\ell}^2+84\eta\widetilde{\ell}^2+24\widetilde{L}^2+6\eta^2\widetilde{L}^2+2\alpha\eta^2\widetilde{\ell}^2)}{\alpha^2} - \frac{64\gamma^2\widetilde{\ell}^2(8\kappa+1)}{\eta} \\
&\qquad - \frac{16\gamma^2\widetilde{L}^2(8\kappa+1)}{\eta^2}\Big) \\
&\overset{(i)}{\geq} 0.
\end{aligned}$$

where (i) are due to $\eta \leq 1$, $\alpha \leq 1$, and the assumption on step-size.

Finally, by using the choice of the momentum parameter

$$\begin{aligned}
\eta \leq \min\Bigg\{&\Big(\frac{40\alpha^2\delta_0\sqrt{(8\kappa+1)\widetilde{\ell}^2}}{96(8\kappa+1)\sigma^2T}\Big)^{2/11}, \Big(\frac{40\alpha^2\delta_0\sqrt{(8\kappa+1)\widetilde{\ell}^2}}{672(8\kappa+1)\sigma^2T}\Big)^{2/7}, \Big(\frac{40\delta_0\sqrt{(8\kappa+1)\widetilde{\ell}^2}}{8(60\kappa+7)\sigma^2T}\Big)^{2/3}, \\
&\Big(\frac{40\delta_0 G\sqrt{(8\kappa+1)\widetilde{\ell}^2}}{8\sigma^2T}\Big)^{2/3}, \Big(\frac{40\alpha\delta_0\sqrt{(8\kappa+1)\widetilde{\ell}^2}}{16(8\kappa+1)\sigma^2T}\Big)^{2/9}\Bigg\},
\end{aligned}$$

ensures that $\frac{96\eta^5(\kappa+1)\sigma^2}{\alpha^2} \leq \frac{40\sqrt{(8\kappa+1)\widetilde{\ell}^2}}{\sqrt{\eta}T}, \frac{672\eta^3(\kappa+1)\sigma^2}{\alpha^2} \leq \frac{40\sqrt{(8\kappa+1)\widetilde{\ell}^2}}{\sqrt{\eta}T}, 8\eta(60\kappa+7)\sigma^2 \leq \frac{40\sqrt{(8\kappa+1)\widetilde{\ell}^2}}{\sqrt{\eta}T}, \frac{8\eta\sigma^2}{G} \leq$

$\frac{40\sqrt{(8\kappa+1)\widetilde{\ell}^2}}{\sqrt{\eta}T}$, and $\frac{16\eta^4(8\kappa+1)\sigma^2}{\alpha} \leq \frac{40\sqrt{(8\kappa+1)\widetilde{\ell}^2}}{\sqrt{\eta}T}$, we obtain

$$
\begin{aligned}
&\mathbb{E}\Big[\|\nabla f(\hat{x}^{(T)})\|^2\Big] \\
&\leq \Big(\frac{40(96(8\kappa+1))^{1/10}\sigma^{2/10}\delta_0\sqrt{(8\kappa+1)\widetilde{\ell}^2}}{\alpha^{2/10}T}\Big)^{10/11} + \Big(\frac{40(672(8\kappa+1))^{1/6}\sigma^{2/6}\delta_0\sqrt{(8\kappa+1)\widetilde{\ell}^2}}{\alpha^{2/6}T}\Big)^{6/7} \\
&+ \Big(\frac{40(8(60\kappa+7))^{1/2}\sigma\delta_0\sqrt{(8\kappa+1)\widetilde{\ell}^2}}{T}\Big)^{2/3} + \Big(\frac{40(\sigma\delta_0\sqrt{8(8\kappa+1)\widetilde{\ell}^2})}{G^{1/2}T}\Big)^{2/3} \\
&+ \Big(\frac{40(16(8\kappa+1))^{1/8}\sigma^{1/4}\delta_0\sqrt{(8\kappa+1)\widetilde{\ell}^2}}{\alpha^{1/8}T}\Big)^{8/9} + 32\kappa\zeta^2 + \frac{\Phi_0}{\gamma T}.
\end{aligned}
$$

This concludes the proof. □

H Missing Proofs for Polyak-Łojasiewicz Functions

In this section, we prove our convergence rates under the (Polyak-Łojasiewicz) PŁ-condition.

H.1 Byz-DM21

Lemma H.1 (Descent lemma)**.** *Suppose that Assumptions 2.1, 2.3, and 3.6 hold. Then for all $s > 0$ we have*

$$\begin{aligned}\mathbb{E}\Big[f(x^{t+1}) - f(x^*)\Big] \leq (1-\gamma\mu)\mathbb{E}\Big[f(x^{(t)}) - f(x^*)\Big] - (\frac{1}{2\gamma} - \frac{L}{2})\mathbb{E}\Big[\|x^{(t+1)} - x^{(t)}\|^2\Big] \\ + \frac{\gamma}{2}\mathbb{E}\Big[\|g^{(t)} - \nabla f(x^{(t)})\|^2\Big].\end{aligned} \tag{38}$$

Proof. The result follows from combining Lemma F.1 with Assumption 3.6. □

Theorem H.2. *Let Assumptions 2.1, 2.3, 2.4, and 3.6 hold. Let us take $\eta \in (0, 1]$ and*

$$0 < \gamma \leq \min\left\{\frac{1}{L + \sqrt{A}}, \frac{\eta}{2\mu}, \frac{\alpha}{4\mu}\right\}$$

where $A = \Big(\frac{104L^2(8\kappa+1)}{\eta^2} + \frac{48L^2(8\kappa+1)(2\eta^4+28\eta^2+\alpha^2+48)}{\alpha^2}\Big)$, Then for all $T \geq 0$ the iterates of Byz-DM21 *satisfy*

$$\begin{aligned}\mathbb{E}\Big[f(x^T) - f(x^*)\Big] \leq &(1-\gamma\mu)^T\mathbb{E}\Big[\Psi^0\Big] + \frac{4\eta\sigma^2(112\kappa G + 13G + 1)}{\mu G} \\ &+ \frac{96\eta^3\sigma^2\Big(8\kappa + 13 + \eta^2(8\kappa+1)\Big)}{\mu\alpha^2} + \frac{16\kappa\zeta^2}{\mu},\end{aligned}$$

where $\mathbb{E}\Big[\Psi^0\Big] = \mathbb{E}\Big[f(x^{(0)}) - f(x^*)\Big] + \frac{8\gamma(8\kappa+1)}{\alpha}\mathbb{E}\Big[\|g^{(0)} - u^{(0)}\|^2\Big] + \frac{4\gamma(8\kappa+1)(\alpha^2+24\eta^2)}{\eta\alpha^2}\mathbb{E}\Big[\|u^{(0)} - v^{(0)}\|^2\Big] + \frac{16\gamma((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2)}{\alpha^2\eta}\mathbb{E}\Big[\|v^{(0)} - \nabla f(x^0)\|^2\Big] + \frac{4\gamma}{\eta}\mathbb{E}\Big[\|\overline{v}^{(0)} - \nabla f(x^0)\|^2\Big]$.

Proof. Let

$$\begin{cases} D, E, F, H \geq 0 \\ D = \frac{8\gamma(8\kappa+1)}{\alpha} \\ E = \frac{4\gamma(8\kappa+1)(\alpha^2+24\eta^2)}{\eta\alpha^2} \\ F = \frac{16\gamma((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2)}{\alpha^2\eta} \\ H = \frac{4\gamma}{\eta} \end{cases} \tag{39}$$

using inequality (28) and Assumption 3.6, we obtain

$$\begin{aligned}&\mathbb{E}\Big[f(x^{(t+1)}) - f(x^*)\Big] \\ &\leq (1-\gamma\mu)\mathbb{E}\Big[f(x^{(t)}) - f(x^*)\Big] - (\frac{1}{2\gamma} - \frac{L}{2})\mathbb{E}\Big[\|x^{(t+1)} - x^{(t)}\|^2\Big] + \frac{\gamma}{2}\Psi^{(t)} \\ &\leq (1-\gamma\mu)\mathbb{E}\Big[f(x^{(t)}) - f(x^*)\Big] - (\frac{1}{2\gamma} - \frac{L}{2})\mathbb{E}\Big[\|x^{(t+1)} - x^{(t)}\|^2\Big] + 2\gamma\mathbb{E}\Big[\|\widetilde{M}^{(t)}\|^2\Big] + 16\gamma\kappa\zeta^2 \\ &\quad + \frac{2\gamma(8\kappa+1)}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|C_i^{(t)}\|^2\Big] + \frac{2\gamma(8\kappa+1)}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|P_i^{(t)}\|^2\Big] + \frac{16\gamma\kappa}{G}\sum_{i\in\mathcal{G}}\mathbb{E}\Big[\|M^{(t)}\|^2\Big] \\ &\leq (1-\gamma\mu)\mathbb{E}\Big[f(x^{(t)}) - f(x^*)\Big] - (\frac{1}{2\gamma} - \frac{L}{2})\mathbb{E}\Big[\|x^{(t+1)} - x^{(t)}\|^2\Big] + 2\gamma\mathbb{E}\Big[\|\widetilde{M}^{(t)}\|^2\Big] + 16\gamma\kappa\zeta^2 \\ &\quad + 2\gamma(8\kappa+1)\mathbb{E}\Big[\|C^{(t)}\|^2\Big] + 2\gamma(8\kappa+1)\mathbb{E}\Big[\|P^{(t)}\|^2\Big] + 16\gamma\kappa\mathbb{E}\Big[\|M^{(t)}\|^2\Big].\end{aligned} \tag{40}$$

Let $\delta^{(t+1)} \overset{\text{def}}{=} f(x^{(t+1)}) - f(x^*)$, $R^{(t)} \overset{\text{def}}{=} x^{(t+1)} - x^{(t)}$. By adding $D \cdot \mathbb{E}\Big[\|C^{(t+1)}\|^2\Big]$ and using Lemma F.3, we have

$$
\begin{aligned}
&\mathbb{E}\Big[\delta^{(t+1)}\Big] + D \cdot \mathbb{E}\Big[\|C^{(t+1)}\|^2\Big] \\
&\leq (1-\gamma\mu)\mathbb{E}\Big[\delta^{(t)}\Big] - (\frac{1}{2\gamma} - \frac{L}{2})\mathbb{E}\Big[\|R^{(t)}\|^2\Big] + 2\gamma\mathbb{E}\Big[\|\widetilde{M}^{(t)}\|^2\Big] + 16\gamma\kappa\zeta^2 \\
&\quad + 2\gamma(8\kappa+1)\mathbb{E}\Big[\|C^{(t)}\|^2\Big] + 2\gamma(8\kappa+1)\mathbb{E}\Big[\|P^{(t)}\|^2\Big] + 16\gamma\kappa\mathbb{E}\Big[\|M^{(t)}\|^2\Big] \\
&\quad + D \cdot \Bigg((1-\frac{\alpha}{2})\mathbb{E}\left[\|C^{(t)}\|^2\right] + \frac{6\eta^4}{\alpha}\mathbb{E}\left[\|M^{(t)}\|^2\right] + \eta^4\sigma^2 + \frac{6\eta^4L^2}{\alpha}\mathbb{E}\left[\|R^{(t)}\|^2\right] \\
&\quad + \frac{6\eta^2}{\alpha}\mathbb{E}\left[\|P^{(t)}\|^2\right]\Bigg) \\
&\leq (1-\gamma\mu)\mathbb{E}\Big[\delta^{(t)}\Big] - (\frac{1}{2\gamma} - \frac{L}{2})\mathbb{E}\Big[\|R^{(t)}\|^2\Big] + 2\gamma\mathbb{E}\Big[\|\widetilde{M}^{(t)}\|^2\Big] + 16\gamma\kappa\zeta^2 \\
&\quad + 2\gamma(8\kappa+1)\mathbb{E}\Big[\|C^{(t)}\|^2\Big] + 2\gamma(8\kappa+1)\mathbb{E}\Big[\|P^{(t)}\|^2\Big] + 16\gamma\kappa\mathbb{E}\Big[\|M^{(t)}\|^2\Big] \\
&\quad + D \cdot (1-\frac{\alpha}{2})\mathbb{E}\left[\|C^{(t)}\|^2\right] + D \cdot \frac{6\eta^4}{\alpha}\mathbb{E}\left[\|M^{(t)}\|^2\right] + D \cdot \eta^4\sigma^2 + D \cdot \frac{6\eta^4L^2}{\alpha}\mathbb{E}\left[\|R^{(t)}\|^2\right] \\
&\quad + D \cdot \frac{6\eta^2}{\alpha}\mathbb{E}\left[\|P^{(t)}\|^2\right].
\end{aligned}
$$

Using inequality (39) and substituting $D = \frac{8\gamma(8\kappa+1)}{\alpha}$, we get

$$
\begin{aligned}
&\mathbb{E}\Big[\delta^{(t+1)}\Big] + D \cdot \mathbb{E}\Big[\|C^{(t+1)}\|^2\Big] \\
&\leq (1-\gamma\mu)\mathbb{E}\Big[\delta^{(t)}\Big] - (\frac{1}{2\gamma} - \frac{L}{2} - \frac{48\gamma\eta^4L^2(8\kappa+1)}{\alpha^2})\mathbb{E}\Big[\|R^{(t)}\|^2\Big] + 2\gamma\mathbb{E}\Big[\|\widetilde{M}^{(t)}\|^2\Big] + 16\gamma\kappa\zeta^2 \\
&\quad + D \cdot (1-\frac{\alpha}{4})\mathbb{E}\left[\|C^{(t)}\|^2\right] + \frac{2\gamma(8\kappa+1)(\alpha^2+24\eta^2)}{\alpha^2}\mathbb{E}\left[\|P^{(t)}\|^2\right] \\
&\quad + \frac{16\gamma(3\eta^4(8\kappa+1)+\kappa\alpha^2)}{\alpha^2}\mathbb{E}\left[\|M^{(t)}\|^2\right] + \frac{8\gamma\eta^4\sigma^2(8\kappa+1)}{\alpha}.
\end{aligned}
$$

Next, by adding $E \cdot \mathbb{E}\Big[\|P^{(t+1)}\|^2\Big]$ and using Lemma F.4, we have

$$
\begin{aligned}
&\mathbb{E}\Big[\delta^{(t+1)}\Big] + D \cdot \mathbb{E}\Big[\|C^{(t+1)}\|^2\Big] + E \cdot \mathbb{E}\left[\|P^{(t+1)}\|^2\right] \\
&\leq (1-\gamma\mu)\mathbb{E}\Big[\delta^{(t)}\Big] - (\frac{1}{2\gamma} - \frac{L}{2} - \frac{48\gamma\eta^4L^2(8\kappa+1)}{\alpha^2})\mathbb{E}\Big[\|R^{(t)}\|^2\Big] + 2\gamma\mathbb{E}\Big[\|\widetilde{M}^{(t)}\|^2\Big] + 16\gamma\kappa\zeta^2 \\
&\quad + \frac{8\gamma\eta^4\sigma^2(8\kappa+1)}{\alpha} + \frac{16\gamma(3\eta^4(8\kappa+1)+\kappa\alpha^2)}{\alpha^2}\mathbb{E}\left[\|M^{(t)}\|^2\right] \\
&\quad + D \cdot (1-\frac{\alpha}{4})\mathbb{E}\left[\|C^{(t)}\|^2\right] + \frac{2\gamma(8\kappa+1)(\alpha^2+24\eta^2)}{\alpha^2}\mathbb{E}\left[\|P^{(t)}\|^2\right] \\
&\quad + E \cdot \Big((1-\eta)\mathbb{E}\Big[\|P^{(t)}\|^2\Big] + 6\eta\mathbb{E}\Big[\|M^{(t)}\|^2\Big] + 6\eta L^2\mathbb{E}\Big[\|R^{(t)}\|^2\Big] + \eta^2\sigma^2\Big).
\end{aligned}
$$

Using inequality (39) and substituting $E = \frac{4\gamma(8\kappa+1)(\alpha^2+24\eta^2)}{\eta\alpha^2}$, we attain

$$
\begin{aligned}
&\mathbb{E}\Big[\delta^{(t+1)}\Big] + D\cdot\mathbb{E}\Big[\|C^{(t+1)}\|^2\Big] + E\cdot\mathbb{E}\left[\|P^{(t+1)}\|^2\right] \\
&\leq (1-\gamma\mu)\mathbb{E}\Big[\delta^{(t)}\Big] - \Big(\frac{1}{2\gamma} - \frac{L}{2} - \frac{48\gamma\eta^4L^2(8\kappa+1)}{\alpha^2} - \frac{24\gamma L^2(8\kappa+1)(\alpha^2+24\eta^2)}{\alpha^2}\Big)\mathbb{E}\Big[\|R^{(t)}\|^2\Big] \\
&\quad + 2\gamma\mathbb{E}\Big[\|\widetilde{M}^{(t)}\|^2\Big] + \frac{4\eta\gamma\sigma^2(8\kappa+1)(\alpha^2+24\eta^2)}{\alpha^2} + 16\gamma\kappa\zeta^2 \\
&\quad + D\cdot(1-\frac{\alpha}{4})\mathbb{E}\left[\|C^{(t)}\|^2\right] + E\cdot(1-\frac{\eta}{2})\mathbb{E}\left[\|P^{(t)}\|^2\right] \\
&\quad + \frac{8\gamma\Big((8\kappa+1)(3\alpha^2+72\eta^2) + 6\eta^4(8\kappa+1) + 2\kappa\alpha^2\Big)}{\alpha^2}\mathbb{E}\Big[\|M^{(t)}\|^2\Big] + \frac{8\gamma\eta^4\sigma^2(8\kappa+1)}{\alpha}.
\end{aligned}
$$

Then, by adding $F\cdot\mathbb{E}\Big[\|M^{(t+1)}\|^2\Big]$, using Lemma F.5, and substituting $F = \frac{16\gamma((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2)}{\alpha^2\eta}$, we obtain

$$
\begin{aligned}
&\mathbb{E}\Big[\delta^{(t+1)}\Big] + D\cdot\mathbb{E}\Big[\|C^{(t+1)}\|^2\Big] + E\cdot\mathbb{E}\left[\|P^{(t+1)}\|^2\right] + F\cdot\mathbb{E}\Big[\|M^{(t+1)}\|^2\Big] \\
&\leq (1-\gamma\mu)\mathbb{E}\Big[\delta^{(t)}\Big] - \Big(\frac{1}{2\gamma} - \frac{L}{2} - \frac{48\gamma\eta^4L^2(8\kappa+1)}{\alpha^2} - \frac{24\gamma L^2(8\kappa+1)(\alpha^2+24\eta^2)}{\alpha^2}\Big)\mathbb{E}\Big[\|R^{(t)}\|^2\Big] \\
&\quad + D\cdot(1-\frac{\alpha}{4})\mathbb{E}\left[\|C^{(t)}\|^2\right] + E\cdot(1-\frac{\eta}{2})\mathbb{E}\left[\|P^{(t)}\|^2\right] + \frac{4\eta\gamma\sigma^2(8\kappa+1)(\alpha^2+24\eta^2)}{\alpha^2} + 16\gamma\kappa\zeta^2 \\
&\quad + \frac{8\gamma\Big((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4) + 2\kappa\alpha^2\Big)}{\alpha^2}\mathbb{E}\Big[\|M^{(t)}\|^2\Big] + \frac{8\gamma\eta^4\sigma^2(8\kappa+1)}{\alpha} + 2\gamma\mathbb{E}\Big[\|\widetilde{M}^{(t)}\|^2\Big] \\
&\quad + F\cdot\Big(\mathbb{E}\left[\|M^{(t)}\|^2\right] + \frac{L^2}{\eta}\mathbb{E}\left[\|R^{(t)}\|^2\right] + \eta^2\sigma^2\Big) \\
&\leq (1-\gamma\mu)\mathbb{E}\Big[\delta^{(t)}\Big] + D\cdot(1-\frac{\alpha}{4})\mathbb{E}\left[\|C^{(t)}\|^2\right] + E\cdot(1-\frac{\eta}{2})\mathbb{E}\left[\|P^{(t)}\|^2\right] \\
&\quad + \frac{4\eta\gamma\sigma^2(8\kappa+1)(\alpha^2+24\eta^2+2\alpha\eta^3)}{\alpha^2} + F\cdot(1-\frac{\eta}{2})\mathbb{E}\Big[\|M^{(t)}\|^2\Big] \\
&\quad - \Bigg(\frac{1}{2\gamma} - \frac{L}{2} - \frac{24\gamma L^2(8\kappa+1)(2\eta^4+\alpha^2+24\eta^2)}{\alpha^2} \\
&\qquad - \frac{16\gamma L^2\Big((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4) + 2\kappa\alpha^2\Big)}{\alpha^2\eta^2}\Bigg)\mathbb{E}\Big[\|R^{(t)}\|^2\Big] \\
&\quad + 2\gamma\mathbb{E}\Big[\|\widetilde{M}^{(t)}\|^2\Big] + 16\gamma\kappa\zeta^2 + \frac{16\gamma\eta\sigma^2\Big((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4) + 2\kappa\alpha^2\Big)}{\alpha^2}.
\end{aligned}
$$

Furthermore, by adding $H \cdot \mathbb{E}\Big[\|\widetilde{M}^{(t+1)}\|^2\Big]$, using Lemma F.5, and substituting $H = \frac{4\gamma}{\eta}$, we arrive at

$$
\begin{aligned}
&\mathbb{E}\Big[\delta^{(t+1)}\Big] + D \cdot \mathbb{E}\Big[\|C^{(t+1)}\|^2\Big] + E \cdot \mathbb{E}\left[\|P^{(t+1)}\|^2\right] + F \cdot \mathbb{E}\Big[\|M^{(t+1)}\|^2\Big] + H \cdot \mathbb{E}\Big[\|\widetilde{M}^{(t+1)}\|^2\Big] \\
&\leq (1-\gamma\mu)\mathbb{E}\Big[\delta^{(t)}\Big] + D \cdot (1-\frac{\alpha}{4})\mathbb{E}\left[\|C^{(t)}\|^2\right] + E \cdot (1-\frac{\eta}{2})\mathbb{E}\left[\|P^{(t)}\|^2\right] \\
&\quad + \frac{4\eta\gamma\sigma^2(8\kappa+1)(\alpha^2+24\eta^2+2\alpha\eta^3)}{\alpha^2} + 16\gamma\kappa\zeta^2 \\
&\quad - \Bigg(\frac{1}{2\gamma} - \frac{L}{2} - \frac{24\gamma L^2(8\kappa+1)(2\eta^4+\alpha^2+24\eta^2)}{\alpha^2} \\
&\qquad - \frac{16\gamma L^2\Big((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2\Big)}{\alpha^2\eta^2}\Bigg)\mathbb{E}\Big[\|R^{(t)}\|^2\Big] \\
&\quad + F \cdot (1-\frac{\eta}{2})\mathbb{E}\Big[\|M^{(t)}\|^2\Big] + \frac{16\gamma\eta\sigma^2\Big((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2\Big)}{\alpha^2} \\
&\quad + 2\gamma\mathbb{E}\Big[\|\widetilde{M}^{(t)}\|^2\Big] + H \cdot \Big((1-\eta)\mathbb{E}\left[\|\widetilde{M}^{(t)}\|^2\right] + \frac{L^2}{\eta}\mathbb{E}\left[\|R^{(t)}\|^2\right] + \frac{\eta^2\sigma^2}{G}\Big) \\
&\leq (1-\gamma\mu)\mathbb{E}\Big[\delta^{(t)}\Big] + D \cdot (1-\frac{\alpha}{4})\mathbb{E}\left[\|C^{(t)}\|^2\right] + E \cdot (1-\frac{\eta}{2})\mathbb{E}\left[\|P^{(t)}\|^2\right] \\
&\quad + \frac{4\eta\gamma\sigma^2(8\kappa+1)(\alpha^2+24\eta^2+2\alpha\eta^3)}{\alpha^2} + 16\gamma\kappa\zeta^2 \\
&\quad + F \cdot (1-\frac{\eta}{2})\mathbb{E}\Big[\|M^{(t)}\|^2\Big] + \frac{16\gamma\eta\sigma^2\Big((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2\Big)}{\alpha^2} \\
&\quad + H \cdot (1-\frac{\eta}{2})\mathbb{E}\left[\|\widetilde{M}^{(t)}\|^2\right] + \frac{4\gamma L^2}{\eta^2}\mathbb{E}\left[\|R^{(t)}\|^2\right] + \frac{4\gamma\eta\sigma^2}{G}.
\end{aligned}
$$

Finally, we bound $R^{(t)}$,

$$
\begin{aligned}
&\Bigg(\frac{1}{2\gamma} - \frac{L}{2} - \frac{4\gamma L^2}{\eta^2} - \frac{24\gamma L^2(8\kappa+1)(2\eta^4+\alpha^2+24\eta^2)}{\alpha^2} \\
&\quad - \frac{16\gamma L^2\Big((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2\Big)}{\alpha^2\eta^2}\Bigg)\mathbb{E}\Big[\|R^{(t)}\|^2\Big] \geq 0.
\end{aligned}
$$

Let

$$
\begin{aligned}
A &\stackrel{\text{def}}{=} \Big(\frac{8L^2}{\eta^2} + \frac{48L^2(8\kappa+1)(2\eta^4+\alpha^2+24\eta^2)}{\alpha^2} + \frac{32L^2\Big((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2\Big)}{\alpha^2\eta^2}\Big) \\
&= \Big(\frac{104L^2(8\kappa+1)}{\eta^2} + \frac{48L^2(8\kappa+1)(2\eta^4+28\eta^2+\alpha^2+48)}{\alpha^2}\Big).
\end{aligned}
$$

Taking $0 < \gamma \leq \frac{1}{L+\sqrt{A}}$ and applying Lemma E.1 gives

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

Let $\mathbb{E}\Big[\Psi^{(t+1)}\Big] \stackrel{\text{def}}{=} \mathbb{E}\Big[\delta^{(t+1)}\Big] + D \cdot \mathbb{E}\Big[\|C^{(t+1)}\|^2\Big] + E \cdot \mathbb{E}\Big[\|P^{(t+1)}\|^2\Big] + F \cdot \mathbb{E}\Big[\|M^{(t+1)}\|^2\Big] + H \cdot \mathbb{E}\Big[\|\widetilde{M}^{(t+1)}\|^2\Big]$ and we use the assumption on η and α to establish that $1 - \frac{\alpha}{4} \leq 1 - \gamma\mu$ and $1 - \frac{\eta}{2} \leq 1 - \gamma\mu$. Applying the inequality

iteratively gives

$$\begin{aligned}\mathbb{E}\Big[\Psi^T\Big] &\leq (1-\gamma\mu)^T\mathbb{E}\Big[\Psi^0\Big] + \frac{4\eta\sigma^2(8\kappa+1)(\alpha^2+24\eta^2+2\alpha\eta^3)}{\mu\alpha^2}\\ &\quad + \frac{16\eta\sigma^2\Big((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2\Big)}{\mu\alpha^2} + \frac{16\kappa\zeta^2}{\mu} + \frac{4\eta\sigma^2}{\mu G}\\ &\leq (1-\gamma\mu)^T\mathbb{E}\Big[\Psi^0\Big] + \frac{4\eta\sigma^2(8\kappa+1)}{\mu} + \frac{96\eta^3\sigma^2(8\kappa+1)}{\mu\alpha^2} + \frac{8\eta^4\sigma^2(8\kappa+1)}{\mu\alpha}\\ &\quad + \frac{32\eta\sigma^2\kappa}{\mu} + \frac{16\kappa\zeta^2}{\mu} + \frac{4\eta\sigma^2}{\mu G}\\ &\quad + \frac{48\eta\sigma^2(8\kappa+1)}{\mu} + \frac{96\eta\sigma^2(8\kappa+1)(12\eta^2+\eta^4)}{\mu\alpha^2}\\ &\leq (1-\gamma\mu)^T\mathbb{E}\Big[\Psi^0\Big] + \frac{4\eta\sigma^2(112\kappa G+13G+1)}{\mu G}\\ &\quad + \frac{8\eta^3\sigma^2(8\kappa+1)(12\eta^2+\eta\alpha+156)}{\mu\alpha^2} + \frac{16\kappa\zeta^2}{\mu}.\end{aligned}$$

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

Corollary H.3. *Suppose that assumptions from Theorem H.2 hold, momentum* $\eta \leq \min\Big\{\frac{\mu\varepsilon G}{(G(\kappa+1)+1)\sigma^2}, \frac{\mu\alpha^2\varepsilon^{1/3}}{(\kappa+1)\sigma^2}, \frac{\mu\alpha\varepsilon^{1/4}}{(\kappa+1)\sigma^2}\Big\}$*, and for all* $i \in \mathcal{G}$*, then Algorithm 1 needs*

$$T = \widetilde{\mathcal{O}}\left(\frac{(G(\kappa+1)+1)\sigma^2}{\mu\varepsilon G} + \frac{(\kappa+1)\sigma^2}{\mu\alpha^2\varepsilon^{1/3}} + \frac{(\kappa+1)\sigma^2}{\mu\alpha\varepsilon^{1/4}} + \frac{L}{\mu} + \frac{L\sigma^2(G(\kappa+1)+1)\sqrt{(\kappa+1)}}{\mu^2\varepsilon G}\right) \tag{41}$$

communication rounds to get an ε*-solution.*

Proof. Considering the choice of η, we have $\frac{1}{\mu}\Big(\frac{4\eta(112\kappa G+13G+1)}{G} + \frac{8\eta^3(8\kappa+1)(112\eta^2+\eta\alpha+156)}{\alpha^2}\Big)\sigma^2 = \mathcal{O}(\varepsilon)$, which guarantees that $\mathbb{E}\Big[f(x^{(T)})-f(x^*)\Big] \leq \varepsilon$ for $\varepsilon \geq \frac{32\kappa\zeta^2}{\mu}$. Therefore, is it sufficient to take the number of communication rounds equal (41) to get an ε-solution. □

H.2 Byz-VR-DM21

Theorem H.4. *Let Assumptions 2.1, 2.3, 2.4, and 3.6 hold. Let us take* $\eta \in (0,1]$ *and*

$$0 < \gamma \leq \min\left\{\frac{1}{L+\sqrt{A}}, \frac{\eta}{2\mu}, \frac{\alpha}{4\mu}\right\},$$

where $A = 32\Big(\frac{8(8\kappa+1)(L^2+7\eta\ell^2)}{\eta^2} + \frac{(8\kappa+1)(L^2(3\eta^2+24)+\ell^2(12\eta^3+\alpha\eta^2+156\eta))}{\alpha^2}\Big)$*, Then for all* $T \geq 0$ *the iterates of* Byz-VR-DM21 *satisfy*

$$\begin{aligned}\mathbb{E}\Big[f(x^T)-f(x^*)\Big] \leq &(1-\gamma\mu)^T\mathbb{E}\Big[\Psi^0\Big] + \frac{8\eta^3\sigma^2(8\kappa+1)(8\eta^2+2\alpha\eta+56)}{\mu\alpha^2}\\ &+ \frac{4\eta\sigma^2(64\kappa G+7G+1)}{\mu G} + \frac{16\kappa\zeta^2}{\mu},\end{aligned}$$

where $\mathbb{E}\Big[\Psi^0\Big] = \mathbb{E}\Big[f(x^{(0)})-f(x^*)\Big] + \frac{8\gamma(8\kappa+1)}{\alpha}\mathbb{E}\Big[\|g^{(0)}-u^{(0)}\|^2\Big] + \frac{4\gamma(8\kappa+1)(\alpha^2+24\eta^2)}{\eta\alpha^2}\mathbb{E}\Big[\|u^{(0)}-v^{(0)}\|^2\Big] + \frac{16\gamma((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2)}{\alpha^2\eta}\mathbb{E}\Big[\|v^{(0)}-\nabla f(x^0)\|^2\Big] + \frac{4\gamma}{\eta}\mathbb{E}\Big[\|\overline{v}^{(0)}-\nabla f(x^0)\|^2\Big]$.

Proof. Let

$$\begin{cases} D, E, F, H \geq 0 \\ D = \frac{8\gamma(8\kappa+1)}{\alpha} \\ E = \frac{4\gamma(8\kappa+1)(\alpha^2+24\eta^2)}{\eta\alpha^2} \\ F = \frac{16\gamma((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2)}{\alpha^2\eta} \\ H = \frac{4\gamma}{\eta} \end{cases} \tag{42}$$

Using inequality (40), we obtain

$$\begin{aligned} &\mathbb{E}\Big[f(x^{(t+1)}) - f(x^*)\Big] \\ &\leq (1-\gamma\mu)\mathbb{E}\Big[f(x^{(t)}) - f(x^*)\Big] - (\frac{1}{2\gamma} - \frac{L}{2})\mathbb{E}\Big[\|x^{(t+1)} - x^{(t)}\|^2\Big] + 2\gamma\mathbb{E}\Big[\|\widetilde{M}^{(t)}\|^2\Big] + 16\gamma\kappa\zeta^2 \\ &\quad + 2\gamma(8\kappa+1)\mathbb{E}\Big[\|C^{(t)}\|^2\Big] + 2\gamma(8\kappa+1)\mathbb{E}\Big[\|P^{(t)}\|^2\Big] + 16\gamma\kappa\mathbb{E}\Big[\|M^{(t)}\|^2\Big]. \end{aligned} \tag{43}$$

Let $\delta^{(t+1)} \stackrel{\text{def}}{=} f(x^{(t+1)}) - f(x^*)$, $R^{(t)} \stackrel{\text{def}}{=} x^{(t+1)} - x^{(t)}$ and by adding $D \cdot \mathbb{E}\Big[\|C^{(t+1)}\|^2\Big]$, by using Lemma G.2, and substituting $D = \frac{8\gamma(8\kappa+1)}{\alpha}$, we have

$$\begin{aligned} &\mathbb{E}\Big[\delta^{(t+1)}\Big] + D \cdot \mathbb{E}\Big[\|C^{(t+1)}\|^2\Big] \\ &\leq (1-\gamma\mu)\mathbb{E}\Big[\delta^{(t)}\Big] - (\frac{1}{2\gamma} - \frac{L}{2})\mathbb{E}\Big[\|R^{(t)}\|^2\Big] + 2\gamma\mathbb{E}\Big[\|\widetilde{M}^{(t)}\|^2\Big] + 16\gamma\kappa\zeta^2 \\ &\quad + 2\gamma(8\kappa+1)\mathbb{E}\Big[\|C^{(t)}\|^2\Big] + 2\gamma(8\kappa+1)\mathbb{E}\Big[\|P^{(t)}\|^2\Big] + 16\gamma\kappa\mathbb{E}\Big[\|M^{(t)}\|^2\Big] \\ &\quad + D \cdot \Bigg((1-\frac{\alpha}{2})\mathbb{E}\left[\|C^{(t)}\|^2\right] + \frac{6\eta^4}{\alpha}\mathbb{E}\left[\|M^{(t)}\|^2\right] + 2\eta^4\sigma^2 + 2\eta^2(\frac{3}{\alpha}L^2 + \ell^2)\mathbb{E}\left[\|R^{(t)}\|^2\right] \\ &\quad + \frac{6\eta^2}{\alpha}\mathbb{E}\left[\|P^{(t)}\|^2\right]\Bigg) \\ &\leq (1-\gamma\mu)\mathbb{E}\Big[\delta^{(t)}\Big] - (\frac{1}{2\gamma} - \frac{L}{2} - \frac{16\gamma\eta^2(8\kappa+1)(\frac{3}{\alpha}L^2+\ell^2)}{\alpha})\mathbb{E}\Big[\|R^{(t)}\Big]\|^2 + 2\gamma\mathbb{E}\Big[\|\widetilde{M}^{(t)}\|^2\Big] \\ &\quad + 16\gamma\kappa\zeta^2 + D \cdot (1-\frac{\alpha}{4})\mathbb{E}\left[\|C^{(t)}\|^2\right] + \frac{2\gamma(8\kappa+1)(\alpha^2+24\eta^2)}{\alpha^2}\mathbb{E}\left[\|P^{(t)}\|^2\right] \\ &\quad + \frac{16\gamma(3\eta^4(8\kappa+1)+\kappa\alpha^2)}{\alpha^2}\mathbb{E}\left[\|M^{(t)}\|^2\right] + \frac{16\gamma\eta^4\sigma^2(8\kappa+1)}{\alpha}. \end{aligned}$$

Next, by adding $E \cdot \mathbb{E}\Big[\|P^{(t+1)}\|^2\Big]$, using Lemma G.3, and substituting $E = \frac{4\gamma(8\kappa+1)(\alpha^2+24\eta^2)}{\eta\alpha^2}$, we get

$$
\begin{aligned}
&\mathbb{E}\Big[\delta^{(t+1)}\Big] + D \cdot \mathbb{E}\Big[\|C^{(t+1)}\|^2\Big] + E \cdot \mathbb{E}\left[\|P^{(t+1)}\|^2\right] \\
&\leq (1-\gamma\mu)\mathbb{E}\Big[\delta^{(t)}\Big] - (\frac{1}{2\gamma} - \frac{L}{2} - \frac{16\gamma\eta^2(8\kappa+1)(\frac{3}{\alpha}L^2+\ell^2)}{\alpha})\mathbb{E}\Big[\|R^{(t)}\|^2\Big] + 2\gamma\mathbb{E}\Big[\|\widetilde{M}^{(t)}\|^2\Big] \\
&\quad + 16\gamma\kappa\zeta^2 + \frac{16\gamma\eta^4\sigma^2(8\kappa+1)}{\alpha} + \frac{16\gamma(3\eta^4(8\kappa+1)+\kappa\alpha^2)}{\alpha^2}\mathbb{E}\left[\|M^{(t)}\|^2\right] \\
&\quad + D \cdot (1-\frac{\alpha}{4})\mathbb{E}\left[\|C^{(t)}\|^2\right] + \frac{2\gamma(8\kappa+1)(\alpha^2+24\eta^2)}{\alpha^2}\mathbb{E}\left[\|P^{(t)}\|^2\right] \\
&\quad + E \cdot \Big((1-\eta)\mathbb{E}\Big[\|P^{(t)}\|^2\Big] + 6\eta\mathbb{E}\Big[\|M^{(t)}\|^2\Big] + 2(\frac{2}{\eta}L^2+\ell^2)\mathbb{E}\Big[\|R^{(t)}\|^2\Big] + 2\eta^2\sigma^2\Big) \\
&\leq (1-\gamma\mu)\mathbb{E}\Big[\delta^{(t)}\Big] + 16\gamma\kappa\zeta^2 + 2\gamma\mathbb{E}\Big[\|\widetilde{M}^{(t)}\|^2\Big] \\
&\quad - \Big(\frac{1}{2\gamma} - \frac{L}{2} - \frac{16\gamma\eta^2(8\kappa+1)(\frac{3}{\alpha}L^2+\ell^2)}{\alpha} - \frac{8\gamma(8\kappa+1)(\frac{2}{\eta}L^2+\ell^2)(\alpha^2+24\eta^2)}{\eta\alpha^2}\Big)\mathbb{E}\Big[\|R^{(t)}\|^2\Big] \\
&\quad + D \cdot (1-\frac{\alpha}{4})\mathbb{E}\left[\|C^{(t)}\|^2\right] + E \cdot (1-\frac{\eta}{2})\mathbb{E}\left[\|P^{(t)}\|^2\right] + \frac{8\eta\gamma\sigma^2(8\kappa+1)(\alpha^2+24\eta^2)}{\alpha^2} \\
&\quad + \frac{8\gamma\Big((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2\Big)}{\alpha^2}\mathbb{E}\Big[\|M^{(t)}\|^2\Big] + \frac{16\gamma\eta^4\sigma^2(8\kappa+1)}{\alpha}.
\end{aligned}
$$

Then, by adding $F \cdot \mathbb{E}\Big[\|M^{(t+1)}\|^2\Big]$, using Lemma G.4, and substituting $F = \frac{16\gamma((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2)}{\alpha^2\eta}$, we obtain

$$
\begin{aligned}
&\mathbb{E}\Big[\delta^{(t+1)}\Big] + D \cdot \mathbb{E}\Big[\|C^{(t+1)}\|^2\Big] + E \cdot \mathbb{E}\left[\|P^{(t+1)}\|^2\right] + F \cdot \mathbb{E}\Big[\|M^{(t+1)}\|^2\Big] \\
&\leq (1-\gamma\mu)\mathbb{E}\Big[\delta^{(t)}\Big] + 16\gamma\kappa\zeta^2 + 2\gamma\mathbb{E}\Big[\|\widetilde{M}^{(t)}\|^2\Big] \\
&\quad - \Big(\frac{1}{2\gamma} - \frac{L}{2} - \frac{16\gamma\eta^2(8\kappa+1)(\frac{3}{\alpha}L^2+\ell^2)}{\alpha} - \frac{8\gamma(8\kappa+1)(\frac{2}{\eta}L^2+\ell^2)(\alpha^2+24\eta^2)}{\eta\alpha^2}\Big)\mathbb{E}\Big[\|R^{(t)}\|^2\Big] \\
&\quad + D \cdot (1-\frac{\alpha}{4})\mathbb{E}\left[\|C^{(t)}\|^2\right] + E \cdot (1-\frac{\eta}{2})\mathbb{E}\left[\|P^{(t)}\|^2\right] + \frac{8\eta\gamma\sigma^2(8\kappa+1)(\alpha^2+24\eta^2)}{\alpha^2} \\
&\quad + \frac{8\gamma\Big((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2\Big)}{\alpha^2}\mathbb{E}\Big[\|M^{(t)}\|^2\Big] + \frac{16\gamma\eta^4\sigma^2(8\kappa+1)}{\alpha} \\
&\quad + F \cdot \Big((1-\eta)\mathbb{E}\left[\|M^{(t)}\|^2\right] + 2\ell^2\mathbb{E}\left[\|R^{(t)}\|^2\right] + 2\eta^2\sigma^2\Big) \\
&\leq (1-\gamma\mu)\mathbb{E}\Big[\delta^{(t)}\Big] + D \cdot (1-\frac{\alpha}{4})\mathbb{E}\left[\|C^{(t)}\|^2\right] + E \cdot (1-\frac{\eta}{2})\mathbb{E}\left[\|P^{(t)}\|^2\right] \\
&\quad + \frac{8\eta\gamma\sigma^2(8\kappa+1)(\alpha^2+24\eta^2+2\alpha\eta^3)}{\alpha^2} + F \cdot (1-\frac{\eta}{2})\mathbb{E}\Big[\|M^{(t)}\|^2\Big] \\
&\quad - \left(\frac{1}{2\gamma} - \frac{L}{2} - \frac{16\gamma\eta^2(8\kappa+1)(\frac{3}{\alpha}L^2+\ell^2)}{\alpha} - \frac{8\gamma(8\kappa+1)(\frac{2}{\eta}L^2+\ell^2)(\alpha^2+24\eta^2)}{\eta\alpha^2}\right. \\
&\quad \left. - \frac{32\gamma\ell^2\Big((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2\Big)}{\eta\alpha^2}\right)\mathbb{E}\Big[\|R^{(t)}\|^2\Big] + 2\gamma\mathbb{E}\Big[\|\widetilde{M}^{(t)}\|^2\Big] \\
&\quad + 16\gamma\kappa\zeta^2 + \frac{32\gamma\eta\sigma^2\Big((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2\Big)}{\alpha^2}.
\end{aligned}
$$

Furthermore, by adding $H \cdot \mathbb{E}\Big[\|\widetilde{M}^{(t+1)}\|^2\Big]$, using Lemma G.4, and substituting $H = \frac{4\gamma}{\eta}$, we arrive at

$$
\begin{aligned}
&\mathbb{E}\Big[\delta^{(t+1)}\Big] + D \cdot \mathbb{E}\Big[\|C^{(t+1)}\|^2\Big] + E \cdot \mathbb{E}\left[\|P^{(t+1)}\|^2\right] + F \cdot \mathbb{E}\Big[\|M^{(t+1)}\|^2\Big] + H \cdot \mathbb{E}\Big[\|\widetilde{M}^{(t+1)}\|^2\Big] \\
&\leq (1-\gamma\mu)\mathbb{E}\Big[\delta^{(t)}\Big] + D \cdot (1-\frac{\alpha}{4})\mathbb{E}\left[\|C^{(t)}\|^2\right] + E \cdot (1-\frac{\eta}{2})\mathbb{E}\left[\|P^{(t)}\|^2\right] + 2\gamma\mathbb{E}\Big[\|\widetilde{M}^{(t)}\|^2\Big] \\
&\quad - \Bigg(\frac{1}{2\gamma} - \frac{L}{2} - \frac{16\gamma\eta^2(8\kappa+1)(\frac{3}{\alpha}L^2+\ell^2)}{\alpha} - \frac{8\gamma(8\kappa+1)(\frac{2}{\eta}L^2+\ell^2)(\alpha^2+24\eta^2)}{\eta\alpha^2} \\
&\quad - \frac{32\gamma\ell^2\Big((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2\Big)}{\eta\alpha^2}\Bigg)\mathbb{E}\Big[\|R^{(t)}\|^2\Big] + F \cdot (1-\frac{\eta}{2})\mathbb{E}\Big[\|M^{(t)}\|^2\Big] + 16\gamma\kappa\zeta^2 \\
&\quad + \frac{32\gamma\eta\sigma^2\Big((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2\Big)}{\alpha^2} + \frac{8\eta\gamma\sigma^2(8\kappa+1)(\alpha^2+24\eta^2+2\alpha\eta^3)}{\alpha^2} \\
&\quad + H \cdot \Big((1-\eta)\mathbb{E}\left[\|\widetilde{M}^{(t)}\|^2\right] + 2\ell^2\mathbb{E}\left[\|R^{(t)}\|^2\right] + \frac{2\eta^2\sigma^2}{G}\Big) \\
&\leq (1-\gamma\mu)\mathbb{E}\Big[\delta^{(t)}\Big] + D \cdot (1-\frac{\alpha}{4})\mathbb{E}\left[\|C^{(t)}\|^2\right] + E \cdot (1-\frac{\eta}{2})\mathbb{E}\left[\|P^{(t)}\|^2\right] + \frac{8\eta\gamma\sigma^2(8\kappa+1)(\alpha^2+24\eta^2+2\alpha\eta^3)}{\alpha^2} \\
&\quad - \Bigg(\frac{1}{2\gamma} - \frac{L}{2} - \frac{16\gamma\eta^2(8\kappa+1)(\frac{3}{\alpha}L^2+\ell^2)}{\alpha} - \frac{8\gamma(8\kappa+1)(\frac{2}{\eta}L^2+\ell^2)(\alpha^2+24\eta^2)}{\eta\alpha^2} - \frac{8\gamma\ell^2}{\eta} \\
&\quad - \frac{32\gamma\ell^2\Big((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2\Big)}{\eta\alpha^2}\Bigg)\mathbb{E}\Big[\|R^{(t)}\|^2\Big] + F \cdot (1-\frac{\eta}{2})\mathbb{E}\Big[\|M^{(t)}\|^2\Big] \\
&\quad + 16\gamma\kappa\zeta^2 + \frac{32\gamma\eta\sigma^2\Big((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2\Big)}{\alpha^2} + H \cdot (1-\frac{\eta}{2})\mathbb{E}\left[\|\widetilde{M}^{(t)}\|^2\right] + \frac{8\gamma\eta\sigma^2}{G}.
\end{aligned}
$$

Finally, we bound $R^{(t)}$,

$$
\begin{aligned}
&\Bigg(\frac{1}{2\gamma} - \frac{L}{2} - \frac{8\gamma\ell^2}{\eta} - \frac{16\gamma\eta^2(8\kappa+1)(\frac{3}{\alpha}L^2+\ell^2)}{\alpha} - \frac{8\gamma(8\kappa+1)(\frac{2}{\eta}L^2+\ell^2)(\alpha^2+24\eta^2)}{\eta\alpha^2} \\
&\quad - \frac{32\gamma\ell^2\Big((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2\Big)}{\eta\alpha^2}\Bigg)\mathbb{E}\Big[\|R^{(t)}\|^2\Big] \geq 0.
\end{aligned}
$$

Let

$$
\begin{aligned}
A &\overset{\text{def}}{=} \Bigg(\frac{16\ell^2}{\eta} + \frac{16(8\kappa+1)(\frac{2}{\eta}L^2+\ell^2)(\alpha^2+24\eta^2)}{\eta\alpha^2} + \frac{64\ell^2\Big((8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)+2\kappa\alpha^2\Big)}{\eta\alpha^2} \\
&\quad + \frac{32\eta^2(8\kappa+1)(\frac{3}{\alpha}L^2+\ell^2)}{\alpha}\Bigg) \\
&= 16\Big(\frac{\ell^2}{\eta} + \frac{(8\kappa+1)(\frac{2}{\eta}L^2+\ell^2)}{\eta} + \frac{24\eta(8\kappa+1)(\frac{2}{\eta}L^2+\ell^2)}{\alpha^2} + \frac{8\kappa\ell^2}{\eta} \\
&\quad + \frac{4\ell^2(8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)}{\eta\alpha^2} + \frac{2\eta^2(8\kappa+1)(\frac{3}{\alpha}L^2+\ell^2)}{\alpha}\Big) \\
&= 16\Big(\frac{2(8\kappa+1)\ell^2}{\eta} + \frac{2(8\kappa+1)L^2}{\eta^2} + \frac{24\eta(8\kappa+1)\ell^2}{\alpha^2} + \frac{48(8\kappa+1)L^2}{\alpha^2} \\
&\quad + \frac{12(8\kappa+1)\ell^2}{\eta} + \frac{4\ell^2(8\kappa+1)(72\eta+6\eta^3)}{\alpha^2} + \frac{6\eta^2(8\kappa+1)L^2}{\alpha^2} + \frac{2\eta^2(8\kappa+1)\ell^2}{\alpha}\Big) \\
&= 16\Big(\frac{(8\kappa+1)(2L^2+14\eta\ell^2)}{\eta^2} + \frac{2(8\kappa+1)((156\eta+12\eta^3+\alpha\eta^2)\ell^2+(3\eta^2+24)L^2)}{\alpha^2}\Big).
\end{aligned}
$$

Taking $0 < \gamma \leq \frac{1}{L+\sqrt{A}}$ and applying Lemma E.1 gives

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

Let $\mathbb{E}\Big[\Psi^{(t+1)}\Big] \overset{\text{def}}{=} \mathbb{E}\Big[\delta^{(t+1)}\Big] + D \cdot \mathbb{E}\Big[\|C^{(t+1)}\|^2\Big] + E \cdot \mathbb{E}\Big[\|P^{(t+1)}\|^2\Big] + F \cdot \mathbb{E}\Big[\|M^{(t+1)}\|^2\Big] + H \cdot \mathbb{E}\Big[\|\widetilde{M}^{(t+1)}\|^2\Big]$ and we use the assumption on η and α to establish that $1 - \frac{\alpha}{4} \leq 1 - \gamma\mu$ and $1 - \frac{\eta}{2} \leq 1 - \gamma\mu$. Applying the inequality iteratively gives

$$\begin{aligned}
&\mathbb{E}\Big[\Psi^T\Big] \\
&\leq (1-\gamma\mu)^T \mathbb{E}\Big[\Psi^0\Big] + \frac{8\eta\sigma^2(8\kappa+1)(\alpha^2+24\eta^2+2\alpha\eta^3)}{\mu\alpha^2} + \frac{32\eta\sigma^2(8\kappa+1)(3\alpha^2+72\eta^2+6\eta^4)}{\mu\alpha^2} \\
&\quad + \frac{64\eta\sigma^2\kappa}{\mu} + \frac{16\kappa\zeta^2}{\mu} + \frac{8\eta\sigma^2}{\mu G} \\
&\leq (1-\gamma\mu)^T \mathbb{E}\Big[\Psi^0\Big] + \frac{8\eta\sigma^2(8\kappa+1)}{\mu} + \frac{16\eta^3\sigma^2(8\kappa+1)(\alpha\eta+12)}{\mu\alpha^2} + \frac{96\eta\sigma^2(8\kappa+1)}{\mu} + \frac{64\eta\sigma^2\kappa}{\mu} \\
&\quad + \frac{192\eta^3\sigma^2(8\kappa+1)(\eta^2+12)}{\mu\alpha^2} + \frac{16\kappa\zeta^2}{\mu} + \frac{8\eta\sigma^2}{\mu G} \\
&\leq (1-\gamma\mu)^T \mathbb{E}[\Psi^0] + \frac{8\eta\sigma^2(108\kappa G+13G+1)}{\mu G} + \frac{16\eta^3\sigma^2(8\kappa+1)(12\eta^2+\alpha\eta+156)}{\mu\alpha^2} + \frac{16\kappa\zeta^2}{\mu}.
\end{aligned}$$

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

Corollary H.5. *Suppose that assumptions from Theorem H.2 hold, momentum* $\eta \leq \min\Big\{\frac{\mu G\varepsilon}{(G(\kappa+1)+1)\sigma^2}, \frac{\mu\alpha^2\varepsilon^{1/3}}{(\kappa+1)\sigma^2}, \frac{\mu\alpha\varepsilon^{1/4}}{(\kappa+1)\sigma^2}\Big\}$*, and for all* $i \in \mathcal{G}$*, then Algorithm 1 needs*

$$T = \widetilde{\mathcal{O}}\left(\frac{(G(\kappa+1)+1)\sigma^2}{\mu\varepsilon G} + \frac{(\kappa+1)\sigma^2}{\mu\alpha^2\varepsilon^{1/3}} + \frac{(\kappa+1)\sigma^2}{\mu\alpha\varepsilon^{1/4}} + \frac{L}{\mu} + \frac{\sigma\sqrt{(\ell^2+L^2)(G(\kappa+1)+1)(\kappa+1)}}{\mu^{3/2}\varepsilon^{1/2}G^{1/2}}\right) \tag{44}$$

communication rounds to get an ε*-solution.*

Proof. Considering the choice of η, we have $\frac{1}{\mu}\Big(\frac{8\eta(108\kappa G+13G+1)}{G} + \frac{16\eta^3(8\kappa+1)(12\eta^2+\alpha\eta+156)}{\alpha^2}\Big)\sigma^2 = \mathcal{O}(\varepsilon)$, which guarantees that $\mathbb{E}\Big[f(x^{(T)}) - f(x^*)\Big] \leq \varepsilon$ for $\varepsilon \geq \frac{32\kappa\zeta^2}{\mu}$. Therefore, is it sufficient to take the number of communication rounds equal (44) to get an ε-solution. □