BYZANTINE ROBUSTNESS AND PARTIAL PARTICIPATION CAN BE ACHIEVED SIMULTANEOUSLY: JUST CLIP GRADIENT DIFFERENCES

Grigory Malinovsky*
AI Initiative
KAUST
Saudi Arabia
grigorii.malinovskii@kaust.edu.sa

Eduard Gorbunov
Department of Machine Learning
MBZUAI
UAE
eduard.gorbunov@mbzuai.ac.ae

Samuel Horváth
Department of Machine Learning
MBZUAI
UAE
samuel.horvath@mbzuai.ac.ae

Peter Richtárik
AI Initiative
KAUST
Saudi Arabia
peter.richtarik@kaust.edu.sa

ABSTRACT

Distributed learning has emerged as a leading paradigm for training large machine learning models. However, in real-world scenarios, participants may be unreliable or malicious, posing a significant challenge to the integrity and accuracy of the trained models. Byzantine fault tolerance mechanisms have been proposed to address these issues, but they often assume full participation from all clients, which is not always practical due to the unavailability of some clients or communication constraints. In our work, we propose the first distributed method with client sampling and provable tolerance to Byzantine workers. The key idea behind the developed method is the use of gradient clipping to control stochastic gradient differences in recursive variance reduction. This allows us to bound the potential harm caused by Byzantine workers, even during iterations when all sampled clients are Byzantine. Furthermore, we incorporate communication compression into the method to enhance communication efficiency. Under general assumptions, we prove convergence rates for the proposed method that match the existing state-of-the-art (SOTA) theoretical results.

1 INTRODUCTION

Distributed optimization problems are a cornerstone of modern machine learning research. They naturally arise in scenarios where data is distributed across multiple clients; for instance, this is typical in Federated Learning (Konečný et al., 2016; Kairouz et al., 2021). Such problems require specialized algorithms adapted to the distributed setup. Additionally, the adoption of distributed optimization methods is motivated by the sheer computational complexity involved in training modern machine learning models. Many models deal with massive datasets and intricate architectures, rendering training infeasible on a single machine (Li, 2020). Distributed methods, by parallelizing computations across multiple machines, offer a pragmatic solution to accelerate training and address these computational challenges, thus pushing the boundaries of machine learning capabilities.

To make distributed training accessible to the broader community, collaborative learning approaches have been actively studied in recent years (Kijsipongse et al., 2018a; Ryabinin & Gusev, 2020; Atre et al., 2021; Diskin et al., 2021a). In such applications, there is a high risk of the occurrence of so-called *Byzantine workers* (Lamport et al., 1982; Su & Vaidya, 2016)—participants who can violate the prescribed distributed algorithm/protocol either intentionally or simply because they are faulty. In general, such workers may even have access to some private data of certain participants and may

*Corresponding author: https://grigory-malinovsky.github.io

collude to increase their impact on the training. Since the ultimate goal is to achieve robustness in the worst case, many papers in the field make no assumptions limiting the power of Byzantine workers. Clearly, in this scenario, standard distributed methods based on the averaging of received information (e.g., stochastic gradients) are not robust, even to a single Byzantine worker. Indeed, such a worker can send an arbitrarily large vector that can shift the method arbitrarily far from the solution. This aspect makes it non-trivial to design distributed methods with provable robustness to Byzantines (Baruch et al., 2019; Xie et al., 2020). Despite all the challenges, multiple methods are developed and analyzed in the literature (Alistarh et al., 2018; Allen-Zhu et al., 2021; Wu et al., 2020; Zhu & Ling, 2021; Karimireddy et al., 2021; 2022; Gorbunov et al., 2022; 2023; Allouah et al., 2023).

However, literally, all existing methods with provable Byzantine robustness require *the full participation of clients*. The requirement of full participation is impractical for modern distributed learning problems since they can have millions of clients (Bonawitz et al., 2017; Niu et al., 2020). In such scenarios, it is more natural to use sampling of clients to speed up the training. Moreover, some clients can be unavailable at certain moments, e.g., due to a poor connection, low battery, or simply because of the need to use the computing power for some other tasks. Although *partial participation of clients* is a natural attribute of large-scale collaborative training, it is not studied under the presence of Byzantine workers. Moreover, this question is highly non-trivial: the existing methods can fail to converge if combined naïvely with partial participation since Byzantine can form a majority during particular rounds and, thus, destroy the whole training with just one round of communications. *Therefore, the field requires the development of new distributed methods that are provably robust to Byzantine attacks and can work with partial participation even when Byzantine workers form a majority during some rounds.*

1.1 Our Contributions

We develop Byzantine-tolerant Variance-Reduced MARINA with Partial Participation (Byz-VR-MARINA-PP, Algorithm 1) – the first distributed method having Byzantine robustness and allowing partial participation of clients. Our method uses variance reduction to handle Byzantine workers and clipping of stochastic gradient differences to bound the potential harm of Byzantine workers even when they form a majority during particular rounds of communication. To make the method even more communication efficient, we add communication compression. We prove the convergence of By-VR-MARINA-PP for general smooth non-convex functions and Polyak-Łojasiewicz functions. In the special case of full participation of clients, our complexity bounds recover the ones for Byz-VR-MARINA (Gorbunov et al., 2023) that are the current SOTA convergence results.

1.2 Related Work

Below we overview closely related works. Additional discussion is deferred to Appendix B.

2 Preliminaries

In this section, we formally introduce the problem, main definition, and assumptions used in the analysis. That is, we consider finite-sum distributed optimization problem[1]

$$\min_{x\in\mathbb{R}^d}\left\{f(x)\overset{\text{def}}{=}\frac{1}{G}\sum_{i\in\mathcal{G}}f_i(x)\right\},\quad f_i(x)\overset{\text{def}}{=}\frac{1}{m}\sum_{j=1}^{m}f_{i,j}(x)\quad\forall i\in\mathcal{G},\tag{1}$$

where $\mathcal{G}$ is a set of regular clients of size $G\overset{\text{def}}{=}|\mathcal{G}|$. In the context of distributed learning, $f_i:\mathbb{R}^d\to\mathbb{R}$ corresponds to the loss function on the data of client i, and $f_{i,j}:\mathbb{R}^d\to\mathbb{R}$ is the loss computed on the j-th sample from the dataset of client i. Next, we assume that the set of all clients taking part in the training is $[n]=\{1,2,\ldots,n\}$ and $\mathcal{G}\subseteq[n]$. The remaining clients $\mathcal{B}\overset{\text{def}}{=}[n]\setminus\mathcal{G}$ are Byzantine ones. We assume that $B\overset{\text{def}}{=}|\mathcal{B}|\leqslant\delta n$, where $0\leqslant\delta<{}^1/_2$ since otherwise Byzantine workers form a majority and problem equation 1 becomes impossible to solve in general.

[1]For simplicity, we assume that all regular workers have the same size of local datasets. Our analysis can be easily generalized to the case of different sizes of local datasets: this will affect only the value of $\mathcal{L}_{\pm}$ from Assumption 6 for some sampling strategies.

Notation. We use a standard notation for the literature on distributed stochastic optimization. Everywhere in the text $\|x\|$ denotes a standard ℓ_2-norm of $x \in \mathbb{R}^d$, $\langle a, b\rangle$ refers to the standard inner product of vectors $a, b \in \mathbb{R}^d$. The clipping operator is defined as follows: $\text{clip}_\lambda(x) \overset{\text{def}}{=} \min\{1, \lambda/\|x\|\}x$ for $x \neq 0$ and $\text{clip}_\lambda(0) \overset{\text{def}}{=} 0$. Finally, $\mathbb{P}\{A\}$ denotes the probability of event A, $\mathbb{E}[\xi]$ is the full expectation of random variable ξ, $\mathbb{E}[\xi \mid A]$ is the expectation of ξ conditioned on the event A. We also sometimes use $\mathbb{E}_k[\xi]$ to denote an expectation of ξ w.r.t. the randomness coming from step k.

Robust aggregator. We follow the definition from (Gorbunov et al., 2023) of (δ, c)-robust aggregation, which is a generalization of the definitions proposed by Karimireddy et al. (2021; 2022).

Definition 2.1 ((δ, c)-Robust Aggregator)**.** *Assume that $\{x_1, x_2, \dots, x_n\}$ is such that there exists a subset $\mathcal{G} \subseteq [n]$ of size $|\mathcal{G}| = G \geqslant (1-\delta)n$ for $\delta \leqslant \delta_{\max} < 0.5$ and there exists $\sigma \geqslant 0$ such that $\frac{1}{G(G-1)}\sum_{i,l\in\mathcal{G}} \mathbb{E}\left[\|x_i - x_l\|^2\right] \leqslant \sigma^2$ where the expectation is taken w.r.t. the randomness of $\{x_i\}_{i\in\mathcal{G}}$. We say that the quantity $\widehat{x}$ is (δ, c)-Robust Aggregator (δ, c)-$\texttt{RAgg}$) and write $\widehat{x} = \texttt{RAgg}(x_1, \dots, x_n)$ for some $c > 0$, if the following inequality holds:*

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

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

One can interpret the definition as follows. Ideally, we would like to filter out all Byzantine workers and compute just an average $\bar{x}$ over the set of good clients. However, this is impossible in general since we do not know apriori who are Byzantine workers. Instead of this, it is natural to expect that the aggregation rule approximates the ideal average up in a certain sense, e.g., in terms of the expected squared distance to $\bar{x}$. As Karimireddy et al. (2021) formally show, in terms of such criterion ($\mathbb{E}[\|\widehat{x} - \bar{x}\|^2]$), the definition of (δ, c)-$\texttt{RAgg}$ cannot be improved (up to the numerical constant). Moreover, standard aggregators such as Krum (Blanchard et al., 2017), geometric median, and coordinate-wise median do not satisfy Definition 2.1 (Karimireddy et al., 2021), though another popular standard aggregation rule called coordinate-wise trimmed mean (Yin et al., 2018) satisfies Definition 2.1 as shown by Allouah et al. (2023) through the more general definition of robust aggregation. To address this issue, Karimireddy et al. (2021) develop the aggregator called CenteredClip and prove that it fits the definition of (δ, c)-$\texttt{RAgg}$. Karimireddy et al. (2022) propose a procedure called Bucketing that fixes Krum, geometric median, and coordinate-wise median, i.e., with Bucketing Krum, geometric, and coordinate-wise median become (δ, c)-$\texttt{ARAgg}$, which is important for our algorithm since the variance of the vectors received from regular workers changes over time in our method. We notice here that δ is a part of the input and can be used when we know apriori that the ratio of Byzantines is smaller than δ; otherwise one can use $\delta = \delta_{\max}$.

Compression operators. In our work, we use standard unbiased compression operators with relatively bounded variance (Khirirat et al., 2018; Horváth et al., 2023).

Definition 2.2 (Unbiased compression)**.** *Stochastic mapping $\mathcal{Q} : \mathbb{R}^d \to \mathbb{R}^d$ is called unbiased compressor/compression operator if there exists $\omega \geqslant 0$ such that for any $x \in \mathbb{R}^d$*

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

For the given unbiased compressor $\mathcal{Q}(x)$, one can define the expected density[2] as $\zeta_{\mathcal{Q}} \overset{def}{=} \sup_{x\in\mathbb{R}^d} \mathbb{E}\left[\|\mathcal{Q}(x)\|_0\right]$, where $\|y\|_0$ is the number of non-zero components of $y \in \mathbb{R}^d$.

In this definition, parameter ω reflects how lossy the compression operator is: the larger ω the more lossy the compression. For example, this class of compression operators includes random sparsification (RandK) (Stich et al., 2018) and quantization (Goodall, 1951; Roberts, 1962; Alistarh et al., 2017). For RandK compression $\omega = \frac{d}{K} - 1, \zeta_{\mathcal{Q}} = K$ and for ℓ_2-quantization $\omega = \sqrt{d} - 1, \zeta_{\mathcal{Q}} = \sqrt{d}$, see the proofs in (Beznosikov et al., 2020).

Assumptions. Up to a couple of assumptions that are specific to our work, we use the same assumptions as in (Gorbunov et al., 2023).

[2]This quantity is well-suited for sparsification-type compression operators like random sparsification (Stich et al., 2018) and 1-level ℓ_2-quantization (Alistarh et al., 2017). For other compressors, such as quantization with more than one level (Goodall, 1951; Roberts, 1962), $\zeta_{\mathcal{Q}}$ is not the main characteristic describing their properties.

Algorithm 1 Byz-VR-MARINA-PP: Byzantine-tolerant VR-MARINA with Partial Participation

Input: starting point x^0, stepsize γ, minibatch size b, probability $p \in (0,1]$, number of iterations K, (δ, c)-ARAgg, clients' sample size $1 \leqslant C \leqslant n$, clipping coefficients $\{\alpha_k\}_{k\geqslant 1}$, direction g^0
for $k = 0, 1, \ldots, K-1$ **do**
  Get a sample from Bernoulli distribution with parameter p: $c_k \sim \mathrm{Be}(p)$
  Sample the set of clients $S_k \subseteq [n]$, $|S_k| = C$ if $c_k = 0$; otherwise $S_k = [n]$
  Broadcast g^k, c_k to all workers
  for $i \in \mathcal{G} \cap S_k$ in parallel **do**
    $x^{k+1} = x^k - \gamma g^k$ and $\lambda_{k+1} = \alpha_{k+1}\|x^{k+1} - x^k\|$
    Set $g_i^{k+1} = \begin{cases} \nabla f_i(x^{k+1}), & \text{if } c_k = 1, \\ g^k + \mathrm{clip}_{\lambda_{k+1}}\left(\mathcal{Q}\left(\widehat{\Delta}_i(x^{k+1}, x^k)\right)\right), & \text{otherwise,} \end{cases}$
    where $\widehat{\Delta}_i(x^{k+1}, x^k)$ is a minibatched estimator of $\nabla f_i(x^{k+1}) - \nabla f_i(x^k)$, $\mathcal{Q}(\cdot)$ for $i \in \mathcal{G} \cap S_k$ are computed independently
  end for
  if $c_k = 1$ **then**
    $g^{k+1} = \mathrm{ARAgg}\left(\{g_i^{k+1}\}_{i\in[n]}\right)$
  else
    $g^{k+1} = g^k + \mathrm{ARAgg}\left(\left\{\mathrm{clip}_{\lambda_{k+1}}\left(\mathcal{Q}\left(\widehat{\Delta}_i(x^{k+1}, x^k)\right)\right)\right\}_{i\in S_k}\right)$
  end if
end for
Return: $\hat{x}^K$ chosen uniformly at random from $\{x^k\}_{k=0}^{K-1}$

We elaborate on additional assumptions in the Appendix.

3 New Method: Byz-VR-MARINA-PP

We propose a new method called Byzantine-tolerant Variance-Reduced MARINA with Partial Participation (Byz-VR-MARINA-PP, Algorithm 1). Our method extends Byz-VR-MARINA (Gorbunov et al., 2023) to the partial participation case via the proper usage of the clipping operator. To illustrate how Byz-VR-MARINA-PP works, we first consider a special case of full participation.

Special case: Byz-VR-MARINA. If all clients participate at each round ($S_k \equiv [n]$) and clipping is turned off ($\lambda_k \equiv +\infty$), then Byz-VR-MARINA-PP reduces to Byz-VR-MARINA that works as follows. Consider the case when no compression is applied ($\mathcal{Q}(x) = x$) and $\widehat{\Delta}_i(x^{k+1}, x^k) = \nabla f_{i,j_k}(x^{k+1}) - \nabla f_{i,j_k}(x^k)$, where j_k is sampled uniformly at random from $[m]$, $i \in \mathcal{G}$. Then, regular workers compute GeomSARAH/PAGE gradient estimator at each step: for $i \in \mathcal{G}$

$$g_i^{k+1} = \begin{cases} \nabla f_i(x^{k+1}), & \text{with probability } p, \\ g^k + \nabla f_{i,j_k}(x^{k+1}) - \nabla f_{i,j_k}(x^k), & \text{otherwise} \end{cases}$$

With small probability p, good workers compute full gradients, and with larger probability $1-p$ they update their estimator via adding stochastic gradient difference. To balance the oracle cost of these two cases, one can choose $p \sim 1/m$ (for minibatched estimator – $p \sim b/m$). Such estimators are known to be optimal for finding stationary points in the stochastic first-order optimization (Fang et al., 2018; Arjevani et al., 2023). Next, good workers send g_i^{k+1} or $\nabla f_{i,j_k}(x^{k+1}) - \nabla f_{i,j_k}(x^k)$ to the server who robustly aggregate the received vectors. Since estimators are conditionally biased, i.e., $\mathbb{E}[g_i^{k+1} \mid x^{k+1}, x^k] \neq \nabla f_i(x^{k+1})$, the additional bias coming from the aggregation does not cause significant issues in the analysis or practice. Moreover, the variance of $\{g_i^{k+1}\}_{i\in\mathcal{G}}$ w.r.t. the sampling of the stochastic gradients is proportional to $\|x^{k+1} - x^k\|^2 \to 0$ with probability $1-p$ (due to Assumption 6) that progressively limits the effect of Byzantine attacks. For a more detailed explanation of why recursive variance reduction works better than SAGA/SVRG-type variance reduction, we refer to (Gorbunov et al., 2023). Arbitrary sampling allows to improve the dependence on the smoothness constants. Unbiased communication compression also naturally fits the framework since it is applied to the stochastic gradient difference, meaning that the variance of $\{g_i^{k+1}\}_{i\in\mathcal{G}}$ w.r.t.

the sampling of the stochastic gradients and compression remains proportional to $\|x^{k+1} - x^k\|^2$ with probability $1 - p$.

New ingredients: client sampling and clipping. The algorithmic novelty of Byz-VR-MARINA-PP in comparison to Byz-VR-MARINA is twofold: with (typically large) probability $1 - p$ only C clients sampled uniformly at random from the set of all clients participate at each round, and clipping is applied to the compressed stochastic gradient differences. With a small probability p, a larger number[3] of clients $\widehat{C} \leqslant n$ takes part in the communication. The main role of clipping is to ensure that the method can withstand the attacks of Byzantines when they form a majority or, more precisely when there are more than $\delta_{\max} C$ Byzantine workers among the sampled ones. *Indeed, without clipping (or some other algorithmic changes) such situations are critical for convergence: Byzantine workers can shift the method arbitrarily far from the solution, e.g., they can collectively send some vector with the arbitrarily large norm.* In contrast, Byz-VR-MARINA-PP tolerates any attacks even when all sampled clients are Byzantine workers since the update remains bounded due to the clipping. Via choosing $\lambda_{k+1} \sim \|x^{k+1} - x^k\|$ we ensure that the norm of transmitted vectors decreases with the same rate as it does in Byz-VR-MARINA with full client participation. Finally, with probability $1 - p$ regular workers can transmit just compressed vectors and leave the clipping operation to the server since Byzantines can ignore clipping operation.

4 Convergence Results

We define $\mathcal{G}_C^k = \mathcal{G} \cap S_k$ and $G_C^k = |\mathcal{G}_C^k|$ and $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ represents the binomial coefficient. We also use the following probabilities:

$$p_G \overset{\text{def}}{=} \mathbb{P}\left\{G_C^k \geqslant (1 - \delta_{\max}) C\right\} = \sum_{\lceil (1-\delta_{\max})C \rceil \leqslant t \leqslant C} \frac{\binom{G}{t}\binom{n-G}{C-t}}{\binom{n}{C}},$$

$$\mathcal{P}_{\mathcal{G}_C^k} \overset{\text{def}}{=} \mathbb{P}\left\{i \in \mathcal{G}_C^k \mid G_C^k \geqslant (1 - \delta_{\max}) C\right\} = \frac{C}{np_G} \cdot \sum_{\lceil (1-\delta_{\max})C \rceil \leqslant t \leqslant C} \frac{\binom{G-1}{t-1}\binom{n-G}{C-t}}{\binom{n-1}{C-1}}.$$

These probabilities naturally appear in the analysis and statements of the theorems. When $c_k = 0$, then server samples C clients, and two situations can appear: either G_C^k is at least $(1 - \delta_{\max}) C$ meaning that the aggregator can ensure robustness according to Definition 2.1 or $G_C^k < (1 - \delta_{\max}) C$. Probability p_G is the probability of the first event, and the second event implies that the aggregation can be spoiled by Byzantine workers (but one can bound the shift using clipping). Finally, we use $\mathcal{P}_{\mathcal{G}_C^k}$ in the computation of some conditional expectations when the first event occurs.

The mentioned probabilities can be easily computed for some special cases. For example, if $C = 1$, then $p_G = {}^G/_n$ and $\mathcal{P}_{\mathcal{G}_C^k} = {}^1/_G$; if $C = 2$, then $p_G = {}^{G(G-1)}/_{n(n-1)}$ and $\mathcal{P}_{\mathcal{G}_C^k} = {}^2/_G$; finally, if $C = n$, then $p_G = 1$ and $\mathcal{P}_{\mathcal{G}_C^k} = 1$.

The next theorem is our main convergence result for general unbiased compression operators.

Theorem 4.1. *Let Assumptions 1, 3, 4, 5, 6 hold and $\lambda_{k+1} = 2\max_{i \in \mathcal{G}} L_i \left\|x^{k+1} - x^k\right\|$. Assume that $0 < \gamma \leqslant {}^1/_{(L+\sqrt{A})}$, where constant A is defined as*

$$A \quad = \quad A_1 L^2 + A_2 \max_{i \in \mathcal{G}} L_i^2 + A_3 L_\pm^2 + A_4 \frac{\mathcal{L}_\pm^2}{b}, \tag{3}$$

[3] As we show next, it is sufficient to take $\widehat{C} \geqslant \frac{\delta n}{\delta_{\max}}$ similarly to (Data & Diggavi, 2021). However, in contrast to the approach from Data & Diggavi (2021), Byz-VR-MARINA-PP requires such communications only with small probability p.

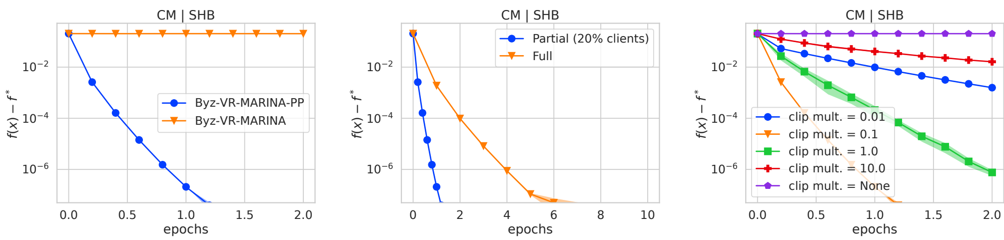

Figure 1: The optimality gap $f(x^k) - f(x^*)$ for 3 different scenarios. We use coordinate-wise mean with bucketing equal to 2 as an aggregation and shift-back as an attack. We use the a9a dataset, where each worker accesses the full dataset with 15 good and 5 Byzantine workers. We do not use any compression. In each step, we sample 20% of clients uniformly at random to participate in the given round unless we specifically mention that we use full participation. Left: Linear convergence of Byz-VR-MARINA-PP with clipping versus non-convergence without clipping. Middle: Full versus partial participation showing faster convergence with clipping. Right: Clipping multiplier λ sensitivity, demonstrating consistent linear convergence across varying λ values.

where

$$
\begin{aligned}
A_1 &= \frac{320}{p^2}\frac{p_G \mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}\omega + \frac{16}{p^2}(1-p_G) + \frac{640}{p^2} p_G \mathcal{P}_{\mathcal{G}_C^k} c\delta_{\max}\omega, \\
A_2 &= \frac{64}{p^2}(1-p_G)F_{\mathcal{A}}^2, \\
A_3 &= \frac{32}{p^2}\frac{p_G \mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}(10\omega+1) + \frac{64}{p^2} p_G \mathcal{P}_{\mathcal{G}_C^k} c\delta_{\max}(10\omega+1), \\
A_4 &= \frac{320}{p^2}\frac{p_G \mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}(\omega+1) + \frac{640}{p^2} p_G \mathcal{P}_{\mathcal{G}_C^k} c\delta_{\max}(\omega+1), \\
\widehat{C} &= 2\frac{\delta_{\max}\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}}{(1-\delta_{\max})}B(6c\delta_{\max}+1).
\end{aligned}
$$

Then for all $K \geqslant 0$ the iterates produced by Byz-VR-MARINA-PP *(Algorithm 1) satisfy*

$$
\mathbb{E}\left[\left\|\nabla f\left(\widehat{x}^K\right)\right\|^2\right] \leqslant \frac{2\Phi^0}{\gamma(K+1)} + \frac{4\widehat{C}\zeta^2}{p}, \tag{4}
$$

where $\widehat{x}^K$ is chosen uniformly at random from $x^0, x^1, \ldots, x^K$, and $\Phi^0 = f\left(x^0\right) - f^* + \frac{2\gamma}{p}\left\|g^0 - \nabla f\left(x^0\right)\right\|^2$. If, in addition, Assumption 8 holds and $0 < \gamma \leqslant 1/(L+\sqrt{2A})$, then for all $K \geqslant 0$ the iterates produced by Byz-VR-MARINA-PP *(Algorithm 1) with $\rho = \min\left\{\gamma\mu, \frac{p}{8}\right\}$ satisfy*

$$
\mathbb{E}\left[f\left(x^K\right) - f\left(x^*\right)\right] \leqslant (1-\rho)^K \Phi^0 + \frac{4\widehat{C}\zeta^2\gamma}{p\rho}, \tag{5}
$$

where $\Phi^0 = f\left(x^0\right) - f^* + \frac{4\gamma}{p}\left\|g^0 - \nabla f\left(x^0\right)\right\|^2$.

The above theorem establishes similar guarantees to the current SOTA ones obtained for Byz-VR-MARINA. That is, in the general non-convex case, we prove $\mathcal{O}(1/K)$ rate, which is optimal (Arjevani et al., 2023), and for PŁ-functions we derive linear convergence result to the neighborhood depending on the heterogeneity. The size of this neighborhood matches the one derived for Byz-VR-MARINA by Gorbunov et al. (2023). It is important to note that our result is obtained considering the scenario of partial participation of clients that results in the more complicated constraints for the stepsize than in (Gorbunov et al., 2023).

Further discussion is available in the appendix.

5 Numerical Experiments

To demonstrate this experimentally, we consider the setup with 15 good workers and 5 Byzantines, *each worker can access the entire dataset*, and the server uses coordinate-wise median with bucketing as the aggregator (see also Appendix E). For the attack, we propose a new attack that we refer to as the *shift-back* attack, which acts in the following way. If Byzantine workers are in the majority in the current round k, then each Byzantine worker sends $x^0 - x^k$. Otherwise, they follow protocol and act as benign workers.

For each experiment, we tune the step size using the following set of candidates $\{0.1, 0.01, 0.001\}$. The step size is fixed. We do not use learning rate warmup or decay. We use batches of size 32 for all methods. For partial participation, in each round, we sample 20% of clients uniformly at random. For $\lambda_k = \lambda\|x^k - x^{k-1}\|$ used for clipping, we select λ from $\{0.1, 1., 10.\}$. Each experiment is run with three varying random seeds, and we report the mean optimality gap with one standard error. The optimal value is obtained by running gradient descent (GD) on the complete dataset for 1000 epochs. Our implementation of attacks and robust aggregation schemes is based on the public implementation from (Gorbunov et al., 2023).

We compare our Byz-VR-MARINA-PP with its version without clipping. We note that the setup that we consider is the most favorable in terms of minimized variance in terms of data and gradient heterogeneity. We show that even in this simplest setup, the method without clipping does not converge since there is no method that can withstand the omniscient Byzantine majority. Therefore, any more complex scenario would also fall short using our simple attack. On the other hand, we show that once clipping is applied, Byz-VR-MARINA-PP is able to converge linearly to the exact solution, complementing our theoretical results.

The main goal of our experimental evaluation is to showcase the benefits of employing clipping to remedy the presence of Byzantine workers and partial participation. For this task, we consider the standard logistic regression model with ℓ_2-regularization, i.e., $f_{i,j}(x) = -y_{i,j}\log(h(x, a_{i,j})) - (1 - y_{i,j})\log(1 - h(x, a_{i,j})) + \eta\|x\|^2$, where $y_{i,j} \in \{0, 1\}$ is the label, $a_{i,j} \in \mathbb{R}^d$ represents the feature vector, η is the regularization parameter, and $h(x, a) = {}^1/_{(1+e^{-a^\top x})}$. This objective is smooth, and for $\lambda > 0$, it is also strongly convex, therefore, it satisfies the PŁ-condition. We consider the *a9a* LIBSVM dataset (Chang & Lin, 2011) and set $\eta = 0.01$. In the experiments, we focus on an important feature of Byz-VR-MARINA-PP: it has linear convergence for homogeneous datasets across clients even in the presence of Byzantine workers and partial participation, as shown in Theorem F.1.

Figure 1 showcases these observations. On the left, we can see Byz-VR-MARINA-PP converges linearly to the optimal solution, while the version without clipping remains stuck at the starting point since Byzantines are always able to push the solution back to the origin since they can create the majority in some rounds. In the middle plot, we compare the full participation scenario in which all the clients participate in each round that does not require clipping since, in each step, we are guaranteed that Byzantines are not in the majority, to partial participation with clipping. We can see, when we compare the total number of computations (measured in epochs), Byz-VR-MARINA-PP leads to faster convergence even though we need to employ clipping. Finally, in the right plot, we measure the sensitivity of clipping multiplier λ. We can see that Byz-VR-MARINA-PP is not very sensitive to λ in terms of convergence, i.e., for all the values of λ, we still converge linearly. However, the suboptimal choice of λ leads to slower convergence.

6 Conclusion and Future Work

This work makes an important first step in the direction of achieving Byzantine robustness under the partial participation of clients. However, some important questions remain open. First of all, it will be interesting to understand whether the derived bounds can be further improved in terms of the dependence on ω, m, and C. Next, one can try to apply the clipping technique to some other Byzantine-robust methods such as SGD with client momentum (Karimireddy et al., 2021; 2022). Finally, the study of other participation patterns (non-uniform sampling/arbitrary client participation) is also a very prominent direction for future research.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.

Ahmad Ajalloeian and Sebastian U Stich. On the convergence of SGD with biased gradients. *arXiv preprint arXiv:2008.00051*, 2020.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pp. 4618–4628, 2018.

Zeyuan Allen-Zhu, Faeze Ebrahimian, Jerry Li, and Dan Alistarh. Byzantine-resilient non-convex stochastic gradient descent. In *International Conference on Learning Representations*, 2021.

Youssef Allouah, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafaël Pinot, and John Stephan. Fixing by mixing: A recipe for optimal byzantine ml under heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pp. 1232–1300, 2023.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2): 165–214, 2023.

Medha Atre, Birendra Jha, and Ashwini Rao. Distributed deep learning using volunteer computing-like paradigm. *arXiv preprint arXiv:2103.08894*, 2021.

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

Debraj Basu, Deepesh Data, Can Karakus, and Suhas Diggavi. Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. signSGD with majority vote is communication efficient and fault tolerant. In *International Conference on Learning Representations*, 2019.

Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *arXiv preprint arXiv:2002.12410*, 2020.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191, 2017.

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

Lingjiao Chen, Hongyi Wang, Zachary Charles, and Dimitris Papailiopoulos. Draco: Byzantine-resilient distributed training via redundant gradients. In *International Conference on Machine Learning*, pp. 903–912, 2018.

Xiangyi Chen, Steven Z Wu, and Mingyi Hong. Understanding gradient clipping in private SGD: A geometric perspective. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13773–13782, 2020.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Georgios Damaskinos, El-Mahdi El-Mhamdi, Rachid Guerraoui, Arsany Guirguis, and Sébastien Rouault. Aggregathor: Byzantine machine learning via robust gradient aggregation. *Proceedings of Machine Learning and Systems*, 1:81–106, 2019.

Deepesh Data and Suhas Diggavi. Byzantine-resilient high-dimensional SGD with local iterations on heterogeneous data. In *International Conference on Machine Learning*, pp. 2478–2488. PMLR, 2021.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Yury Demidovich, Grigory Malinovsky, Igor Sokolov, and Peter Richtárik. A guide through the Zoo of biased SGD. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Michael Diskin, Alexey Bukhtiyarov, Max Ryabinin, Lucile Saulnier, Quentin Lhoest, Anton Sinitsin, Dmitriy Popov, Dmitry Pyrkin, Maxim Kashirin, Alexander Borzunov, Albert Villanova del Moral, Denis Mazur, Ilia Kobelev, Yacine Jernite, Thomas Wolf, and Gennady Pekhimenko. Distributed deep learning in open collaborations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7879–7897, 2021a.

Michael Diskin, Alexey Bukhtiyarov, Max Ryabinin, Lucile Saulnier, Anton Sinitsin, Dmitry Popov, Dmitry V Pyrkin, Maxim Kashirin, Alexander Borzunov, Albert Villanova del Moral, et al. Distributed deep learning in open collaborations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7879–7897, 2021b.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

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

Avishek Ghosh, Raj Kumar Maity, and Arya Mazumdar. Distributed newton can communicate less and resist Byzantine workers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18028–18038, 2020.

Avishek Ghosh, Raj Kumar Maity, Swanand Kadhe, Arya Mazumdar, and Kannan Ramchandran. Communication-efficient and byzantine-robust distributed learning with error feedback. *IEEE Journal on Selected Areas in Information Theory*, 2(3):942–953, 2021.

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

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15042–15053, 2020.

Eduard Gorbunov, Konstantin P Burlachenko, Zhize Li, and Peter Richtárik. MARINA: Faster non-convex distributed learning with compression. In *International Conference on Machine Learning*, pp. 3788–3798, 2021.

Eduard Gorbunov, Alexander Borzunov, Michael Diskin, and Max Ryabinin. Secure distributed training at scale. In *International Conference on Machine Learning*, 2022. (arXiv preprint arXiv:2106.11257, 2021).

Eduard Gorbunov, Samuel Horváth, Peter Richtárik, and Gauthier Gidel. Variance reduction is an antidote to Byzantines: Better rates, weaker assumptions and communication compression as a cherry on the top. *International Conference on Learning Representations*, 2023.

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

Farzin Haddadpour, Mohammad Mahdi Kamani, Aryan Mokhtari, and Mehrdad Mahdavi. Federated learning with compression: Unified analysis and sharp guarantees. In *International Conference on Artificial Intelligence and Statistics*, pp. 2350–2358, 2021.

Lie He, Sai Praneeth Karimireddy, and Martin Jaggi. Byzantine-robust decentralized learning via ClippedGossip. *arXiv preprint arXiv:2202.01545*, 2022.

Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. *Optimization Methods and Software*, 38(1), 2023. (arXiv preprint arXiv:1904.05115, 2019).

Rustem Islamov, Xun Qian, and Peter Richtárik. Distributed second order methods with fast rates and compressed communication. In *International Conference on Machine Learning*, pp. 4617–4628, 2021.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

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

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Learning from history for Byzantine robust optimization. In *International Conference on Machine Learning*, pp. 5311–5319, 2021.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. *International Conference on Learning Representations*, 2022.

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

Ekasit Kijsipongse, Apivadee Piyatumrong, and Suriya U-ruekolan. A hybrid gpu cluster and volunteer computing platform for scalable deep learning. *The Journal of Supercomputing*, 04 2018a. doi: 10.1007/s11227-018-2375-9.

Ekasit Kijsipongse, Apivadee Piyatumrong, et al. A hybrid gpu cluster and volunteer computing platform for scalable deep learning. *The Journal of Supercomputing*, 74(7):3236–3263, 2018b.

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

Leslie Lamport, Robert Shostak, and Marshall Pease. The Byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3):382–401, 1982.

Chuan Li. Demystifying gpt-3 language model: A technical overview, 2020. "https://lambdalabs.com/blog/demystifying-gpt-3".

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

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

Stanislaw Łojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117:87–89, 1963.

Lingjuan Lyu, Han Yu, Xingjun Ma, Lichao Sun, Jun Zhao, Qiang Yang, and Philip S Yu. Privacy and robustness in federated learning: Attacks and defenses. *arXiv preprint arXiv:2012.06337*, 2020.

Vien V Mai and Mikael Johansson. Stability and convergence of stochastic gradient clipping: Beyond Lipschitz continuity and smoothness. In *International Conference on Machine Learning*, pp. 7325–7335, 2021.

Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*, 2019.

Alexander V Nazin, Arkadi S Nemirovsky, Alexandre B Tsybakov, and Anatoli B Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80:1607–1627, 2019.

Ion Necoara, Yurii Nesterov, and François Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175:69–107, 2019.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pp. 2613–2621, 2017.

Ta Duy Nguyen, Alina Ene, and Huy L Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tails. *arXiv preprint arXiv:2304.01119*, 2023.

Chaoyue Niu, Fan Wu, Shaojie Tang, Lifeng Hua, Rongfei Jia, Chengfei Lv, Zhihua Wu, and Guihai Chen. Billion-scale federated learning on mobile clients: A submodel design with tunable privacy. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, pp. 1–14, 2020.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pp. 1310–1318, 2013.

Krishna Pillutla, Sham M Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022.

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

Xun Qian, Peter Richtárik, and Tong Zhang. Error compensated distributed SGD can be accelerated. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Shashank Rajput, Hongyi Wang, Zachary Charles, and Dimitris Papailiopoulos. Detox: A redundancy-based framework for faster and more robust gradient aggregation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jayanth Regatti, Hao Chen, and Abhishek Gupta. ByGARS: Byzantine SGD with arbitrary number of attackers. *arXiv preprint arXiv:2006.13421*, 2020.

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

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

Nuria Rodríguez-Barroso, Eugenio Martínez-Cámara, M Luzón, Gerardo González Seco, Miguel Ángel Veganzones, and Francisco Herrera. Dynamic federated learning model for identifying adversarial clients. *arXiv preprint arXiv:2007.15030*, 2020.

Max Ryabinin and Anton Gusev. Towards crowdsourced training of large neural networks using decentralized mixture-of-experts. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3659–3672. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/25ddc0f8c9d3e22e03d3076f98d83cb2-Paper.pdf.

Abdurakhmon Sadiev, Grigory Malinovsky, Eduard Gorbunov, Igor Sokolov, Ahmed Khaled, Konstantin Burlachenko, and Peter Richtárik. Federated optimization algorithms with random reshuffling and gradient compression. *arXiv preprint arXiv:2206.07021*, 2022.

Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. *arXiv preprint arXiv:2302.00999*, 2023.

Mher Safaryan, Rustem Islamov, Xun Qian, and Peter Richtárik. FedNL: Making Newton-type methods applicable to federated learning. In *International Conference on Machine Learning*, 2022. (arXiv preprint arXiv:2106.02969, 2021).

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, 2017.

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

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Lili Su and Nitin H Vaidya. Fault-tolerant multi-agent optimization: optimal iterative distributed algorithms. In *Proceedings of the 2016 ACM Symposium on Principles of Distributed Computing*, pp. 425–434, 2016.

Rafał Szlendak, Alexander Tyurin, and Peter Richtárik. Permutation compressors for provably faster distributed nonconvex optimization. In *International Conference on Learning Representations*, 2022.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *International Conference on Artificial Intelligence and Statistics*, pp. 1195–1204, 2019.

Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. Powersgd: Practical low-rank gradient compression for distributed optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Zhaoxian Wu, Qing Ling, Tianyi Chen, and Georgios B Giannakis. Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks. *IEEE Transactions on Signal Processing*, 68:4583–4596, 2020.

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

Xinyi Xu and Lingjuan Lyu. Towards building a robust and fair federated learning system. *arXiv preprint arXiv:2011.10464*, 2020.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pp. 5650–5659, 2018.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020a. (arXiv preprint arXiv:1905.11881, 2019).

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems*, volume 33, pp. 15383–15393, 2020b.

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

A APPENDIX

CONTENTS

B Related Work

Byzantine robustness. The primary vulnerability of standard distributed methods to Byzantine attacks lies in the aggregation rule: even one worker can arbitrarily distort the average. Therefore, many papers on Byzantine robustness focus on the application of robust aggregation rules, such as the geometric median (Pillutla et al., 2022), coordinate-wise median, trimmed median (Yin et al., 2018), Krum (Blanchard et al., 2017), and Multi-Krum (Damaskinos et al., 2019). However, simply robustifying the aggregation rule is insufficient to achieve provable Byzantine robustness, as illustrated by Baruch et al. (2019) and Xie et al. (2020), who design special Byzantine attacks that can bypass standard defenses. This implies that more significant algorithmic changes are required to achieve Byzantine robustness, a point also formally proven by Karimireddy et al. (2021), who demonstrate that permutation-invariant algorithms – i.e., algorithms independent of the order of stochastic gradients at each step – cannot provably converge to any predefined accuracy in the presence of Byzantines.

Wu et al. (2020) are the first who exploit variance reduction to tolerate Byzantine attacks. They propose and analyze the method called Byrd-SAGA, which uses SAGA-type (Defazio et al., 2014) gradient estimators on the good workers and geometric median for the aggregation. Gorbunov et al. (2023) develop another variance-reduced method called Byz-VR-MARINA, which is based on (conditionally biased) GeomSARAH/PAGE-type (Horváth et al., 2023; Li et al., 2021) gradient estimator and any robust aggregation in the sense of the definition from (Karimireddy et al., 2021; 2022), and derive the improved convergence guarantees that are the current SOTA in the literature. There are also many other approaches and we discuss some of them in Appendix B.

Partial participation and client sampling. In the context of Byzantine-robust learning, there exists one work that develops and analyzes the method with partial participation (Data & Diggavi, 2021). However, this work relies on the restrictive assumption that the number of participating clients at each round is at least three times larger than the number of Byzantine workers. In this case, Byzantines cannot form a majority, and standard methods can be applied without any changes. In contrast, our method converges in more challenging scenarios, e.g., Byz-VR-MARINA-PP provably converges even when the server samples one client, which can be Byzantine. The results from Data & Diggavi (2021) have some other noticeable limitations that we discuss in Appendix B.

Further Comparison with Data & Diggavi (2021). As we mention in the main text, Data & Diggavi (2021) assume that $3B$ is smaller than C. More precisely, Data & Diggavi (2021) assume that $B \leqslant \epsilon C$, where $\epsilon \leqslant \frac{1}{3} - \epsilon'$ for some parameter $\epsilon' > 0$ that will be explained later. That is, the results from Data & Diggavi (2021) do not hold when C is smaller than $3B$, and, in particular, their algorithm cannot tolerate the situation when the server samples only Byzantine workers at some particular communication round. We also notice that when $C \geqslant 4B$, then existing methods such as Byz-VR-MARINA (Gorbunov et al., 2023) or Client Momentum (Karimireddy et al., 2021; 2022) can be applied without any changes to get a provable convergence.

Next, Data & Diggavi (2021) derive the upper bounds for the expected squared distance to the solution (in the strongly convex case) and the averaged expected squared norm of the gradient (in the non-convex case), where the expectation is taken w.r.t. the sampling of stochastic gradients only and the bounds itself hold with probability at least $1 - \frac{K}{H}\exp\left(-\frac{\epsilon'^2(1-\epsilon)C}{16}\right)$, where H is the number of local steps. For simplicity consider the best-case scenario: $H = 1$ (local steps deteriorate the results from Data & Diggavi (2021)). Then, the lower bound for this probability becomes negative when either C is not large enough or when K is large or when ϵ is close to $\frac{1}{3}$, e.g., for $K = 10^6, \epsilon = \epsilon' = \frac{1}{6}, C = 5000$ this lower bound is smaller than -720, meaning that in this case, the result does not guarantee convergence. In contrast, our results have classical convergence criteria, where the expectations are taken w.r.t. the all randomness.

Finally, the bounds from Data & Diggavi (2021) have non-reduceable terms even for homogeneous data case: these terms are proportional to $\frac{\sigma^2}{b}$, where σ^2 is the upper bound for the variance of the stochastic estimator on regular clients and b is the batchsize. In contrast, our results have only decreasing terms in the upper bounds when the data is homogeneous.

Byzantine robustness. There exist various approaches to achieving Byzantine robustness (Lyu et al., 2020). Alistarh et al. (2018); Allen-Zhu et al. (2021) rely on the concentration inequalities for the stochastic gradients with bounded noise to iteratively remove them from the training. Karimireddy et al. (2021) formalize the definition of robust aggregation and propose the first provably robust aggregation rule called CenteredClip and the first provably Byzantine robust method under bounded variance assumption for homogeneous problems, i.e., when all good workers share one dataset. In particular, the method from (Karimireddy et al., 2021) uses client momentum on the clients that helps to memorize previous steps for good workers and withstand time-coupled attacks. This approach is extended by He et al. (2022) to the setup of decentralized learning. Allouah et al. (2023) develop an alternative definition for robust aggregation and propose a new aggregation rule satisfying their definition. Karimireddy et al. (2022) generalize these results to the heterogeneous data case and derive lower bounds for the optimization error that one can achieve in the heterogeneous case. Based on the formalism from Karimireddy et al. (2021), Gorbunov et al. (2022) propose a server-free approach that uses random checks of computations and bans of peers. This trick allows the elimination of all Byzantine workers after a finite number of steps on average. There are also many other approaches, e.g., one can use redundant computations of the stochastic gradients (Chen et al., 2018; Rajput et al., 2019) or introduce reputation metrics (Rodríguez-Barroso et al., 2020; Regatti et al., 2020; Xu & Lyu, 2020) to achieve some robustness, see also a recent survey by Lyu et al. (2020).

Variance reduction. The literature on variance-reduced methods is very rich (Gower et al., 2020). The first variance-reduced methods are designed to fix the convergence of standard Stochastic Gradient Descent (SGD) and make it convergent to any predefined accuracy even with constant stepsizes. Such methods as SAG (Schmidt et al., 2017), SVRG (Johnson & Zhang, 2013), SAGA (Defazio et al., 2014) are developed mainly for (strongly) convex smooth optimization problems, while methods like SARAH (Nguyen et al., 2017), STORM (Cutkosky & Orabona, 2019), GeomSARAH (Horváth et al., 2023), PAGE (Li et al., 2021) are designed for general smooth non-convex problems. In this paper, we use GeomSARAH/PAGE-type variance reduction as the main building block of the method that makes the method robust to Byzantine attacks.

Partial participation and client sampling. In the context of Byzantine-robust learning, there exists one work that develops and analyzes the method with partial participation (Data & Diggavi, 2021). However, this work relies on the restrictive assumption that the number of participating clients at each round is at least three times larger than the number of Byzantine workers. In this case, Byzantines cannot form a majority, and standard methods can be applied without any changes. In contrast, our method converges in more challenging scenarios, e.g., Byz-VR-MARINA-PP provably converges even when the server samples one client, which can be Byzantine. The results from Data & Diggavi (2021) have some other noticeable limitations that we discuss in Appendix B.

Communication compression. The literature on communication compression can be roughly divided into two huge groups. The first group studies the methods with unbiased communication compression. Different compression operators in the application to Distributed SGD/GD are studied in (Alistarh et al., 2017; Wen et al., 2017; Khirirat et al., 2018). To improve the convergence rate by fixing the error coming from the compression Mishchenko et al. (2019) propose to apply compression to the special gradient differences. Multiple extensions and generalizations of mentioned techniques are proposed and analyzed in the literature, e.g., see (Horváth et al., 2023; Gorbunov et al., 2021; Li et al., 2020; Qian et al., 2021; Basu et al., 2019; Haddadpour et al., 2021; Sadiev et al., 2022; Islamov et al., 2021; Safaryan et al., 2022).

Another large part of the literature on compressed communication is devoted to biased compression operators (Ajalloeian & Stich, 2020; Demidovich et al., 2023). Typically, such compression operators require more algorithmic changes than unbiased compressors since naïve combinations of biased compression with standard methods (e.g., Distributed GD) can diverge (Beznosikov et al., 2020). Error feedback is one of the most popular ways of utilizing biased compression operators in practice (Seide et al., 2014; Stich et al., 2018; Vogels et al., 2019), see also (Richtárik et al., 2021; Fatkhullin et al., 2021) for the modern version of error feedback with better theoretical guarantees for non-convex problems.

In the context of Byzantine robustness, methods with communication compression are also studied. The existing approaches are based on aggregation rules based on the norms of the updates (Ghosh et al., 2020; 2021), SignSGD and majority vote (Bernstein et al., 2019), SAGA-type variance reduction coupled with unbiased compression (Zhu & Ling, 2021), and GeomSARAH/PAGE-type variance reduction combined with unbiased compression (Gorbunov et al., 2023).

Gradient clipping. Gradient clipping has multiple useful properties and applications. Originally it was used by Pascanu et al. (2013) to reduce the effect of exploding gradients during the training of RNNs. Gradient clipping is also a popular tool for achieving provable differential privacy (Abadi et al., 2016; Chen et al., 2020), convergence under generalized notions of smoothness (Zhang et al., 2020a; Mai & Johansson, 2021) and better (high-probability) convergence under heavy-tailed noise assumption (Zhang et al., 2020b; Nazin et al., 2019; Gorbunov et al., 2020; Sadiev et al., 2023; Nguyen et al., 2023). In the context of Byzantine-robust learning, gradient clipping is also utilized to design provably robust aggregation (Karimireddy et al., 2021). Our work proposes a novel useful application of clipping, i.e., we utilize clipping to achieve Byzantine robustness with partial participation of clients.

C USEFUL FACTS

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

$$2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2 \tag{6}$$

$$\|a + b\|^2 \leqslant (1 + \alpha)\|a\|^2 + \left(1 + \alpha^{-1}\right)\|b\|^2 \tag{7}$$

$$-\|a - b\|^2 \leqslant -\frac{1}{1+\alpha}\|a\|^2 + \frac{1}{\alpha}\|b\|^2, \tag{8}$$

$$(1 - p)\left(1 + \frac{p}{2}\right) \leqslant 1 - \frac{p}{2}, \quad p \geqslant 0 \tag{9}$$

$$(1 - p)\left(1 + \frac{p}{2}\right)\left(1 + \frac{p}{4}\right) \leqslant 1 - \frac{p}{4} \quad p \geqslant 0. \tag{10}$$

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

D ASSUMPTIONS

Assumption 1 (Bounded `ARAgg`)**.** *We assume that the server applies aggregation rule $\mathcal{A}$ such that $\mathcal{A}$ is (δ, c)-`ARAgg` and there exists constant $F_{\mathcal{A}} > 0$ such that for any inputs $x_1, \ldots, x_n \in \mathbb{R}^d$ the norm of the aggregator is not greater than the maximal norm of the inputs:*

$$\|\mathcal{A}(x_1, \ldots, x_n)\| \leqslant F_{\mathcal{A}} \max_{i \in [n]} \|x_i\|.$$

The above assumption is satisfied for popular (δ, c)-robust aggregation rules presented in the literature (Karimireddy et al., 2021; 2022). Therefore, this assumption is more a formality than a real limitation: it is needed to exclude some pathological examples of (δ, c)-robust aggregation rules, e.g., for any $\mathcal{A}$ that is (δ, c)-`RAgg` one can construct unbounded $(\delta, 2c)$-`RAgg` as $\overline{\mathcal{A}} = \mathcal{A} + X$, where X is a random sample from the Gaussian distribution $\mathcal{N}(0, c\delta\sigma^2)$.

Next, for part of our results, we also make the following assumption.

Assumption 2 (Bounded compressor (optional))**.** *We assume that workers use compression operator $\mathcal{Q}$ satisfying Definition 2.2 and bounded as follows:*

$$\|\mathcal{Q}(x)\| \leqslant D_Q \|x\| \quad \forall x \in \mathbb{R}^d.$$

For example, RandK and ℓ_2-quantization meet this assumption with $D_{\mathcal{Q}} = \frac{d}{K}$ and $D_{\mathcal{Q}} = \sqrt{d}$ respectively. In general, constant $D_{\mathcal{Q}}$ can be large (proportional to d). However, in practice, one can use RandK with $K = \frac{d}{100}$ and, thus, have moderate $D_{\mathcal{Q}} = 100$. We also have the results without Assumption 2, but with worse dependence on some other parameters, see the discussion in Section 4.

Next, we assume that good workers have ζ^2-heterogeneous local loss functions.

Assumption 3 (ζ^2-heterogeneity)**.** *We assume that good clients have ζ^2-heterogeneous local loss functions for some $\zeta \geqslant 0$, i.e.,*

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

The above assumption is quite standard for the literature on Byzantine robustness (Wu et al., 2020; Karimireddy et al., 2022; Gorbunov et al., 2023; Allouah et al., 2023). Moreover, some kind of a bound on the heterogeneity of good clients is necessary since otherwise Byzantine robustness cannot be achieved in general. In the appendix, all proofs are given under a more general version of Assumption 3, see Assumption 9. Finally, the case of homogeneous data ($\zeta = 0$) is also quite popular for collaborative learning (Diskin et al., 2021b; Kijsipongse et al., 2018b).

The following assumption is classical for the literature on non-convex optimization.

Assumption 4 (L-smoothness)**.** *We assume that function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth, i.e., for all $x, y \in \mathbb{R}^d$ we have $\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|$. Moreover, we assume that f is uniformly lower bounded by $f^* \in \mathbb{R}$, i.e., $f^* \stackrel{def}{=} \inf_{x \in \mathbb{R}^d} f(x)$. In addition, we assume that f_i is L_i-smooth for all $i \in \mathcal{G}$, i.e., for all $x, y \in \mathbb{R}^d$*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leqslant L_i \|x - y\|. \tag{11}$$

We notice here that equation 11 implies L-smoothness of f with $L \leqslant \frac{1}{G}\sum_{i \in \mathcal{G}} L_i$, i.e., smoothness constant of f can be better than the averaged smoothness constant of the local loss functions on the regular clients.

Following Gorbunov et al. (2023), we consider refined assumptions on the smoothness.

Assumption 5 (Global Hessian variance assumption (Szlendak et al., 2022))**.** *We assume that there exists $L_{\pm} \geqslant 0$ such that for all $x, y \in \mathbb{R}^d$*

$$\frac{1}{G} \sum_{i \in \mathcal{G}} \|\nabla f_i(x) - \nabla f_i(y)\|^2 - \|\nabla f(x) - \nabla f(y)\|^2 \leqslant L_{\pm}^2 \|x - y\|^2. \tag{12}$$

Assumption 6 (Local Hessian variance assumption (Gorbunov et al., 2023)). *We assume that there exists $\mathcal{L}_{\pm} \geqslant 0$ such that for all $x, y \in \mathbb{R}^d$*

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

where $\Delta_i(x,y) \stackrel{def}{=} \nabla f_i(x) - \nabla f_i(y)$ and $\widehat{\Delta}_i(x,y)$ is an unbiased mini-batched estimator of $\Delta_i(x,y)$ with batch size b.

We notice that equation 11 implies equation 12 with $L_{\pm} \leqslant \max_{i\in\mathcal{G}} L_i$. Szlendak et al. (2022) prove that $L_{\pm}$ satisfies the following relation: $L_{\text{avg}}^2 - L^2 \leqslant L_{\pm}^2 \leqslant L_{\text{avg}}^2$, where $L_{\text{avg}}^2 \stackrel{\text{def}}{=} \frac{1}{G}\sum_{i\in\mathcal{G}} L_i^2$. In particular, it is possible that $L_{\pm} = 0$ even if the data on the good workers is heterogeneous.

This assumption incorporates considerations for the smoothness characteristics inherent in all functions $\{f_{i,j}\}_{i\in\mathcal{G},j\in[m]}$, the sampling policy, and the similarity among the functions $\{f_{i,j}\}_{i\in\mathcal{G},j\in[m]}$. Gorbunov et al. (2023) have demonstrated that, assuming smoothness of $\{f_{i,j}\}_{i\in\mathcal{G},j\in[m]}$, Assumption 6 holds for various standard sampling strategies, including uniform and importance samplings.

For part of our results, we also need to assume smoothness of all $\{f_{i,j}\}_{i\in\mathcal{G},j\in[m]}$ explicitly.

Assumption 7 (Smoothness of $f_{i,j}$ (optional)). *We assume that for all $i \in \mathcal{G}$ and $j \in [m]$ there exists $L_{i,j} \geqslant 0$ such that $f_{i,j}$ is $L_{i,j}$-smooth, i.e., for all $x, y \in \mathbb{R}^d$*

$$\|\nabla f_{i,j}(x) - \nabla f_{i,j}(y)\| \leqslant L_{i,j}\|x-y\|. \tag{13}$$

Finally, we also consider functions satisfying Polyak-Łojasiewicz (PŁ) condition (Polyak, 1963; Łojasiewicz, 1963). This assumption belongs to the class of assumptions on the structured non-convexity that allows achieving linear convergence (Necoara et al., 2019).

Assumption 8 (PŁ condition (optional)). *We assume that function f satisfies Polyak-Łojasiewicz (PŁ) condition with parameter $\mu > 0$, i.e., for all $x \in \mathbb{R}^d$ there exists $f^* \stackrel{def}{=} \inf_{x\in\mathbb{R}^d} f(x)$ such that*

$$\|\nabla f(x)\|^2 \geqslant 2\mu\left(f(x) - f^*\right).$$

E JUSTIFICATION OF ASSUMPTION 1

Algorithm 2 Bucketing Algorithm (Karimireddy et al., 2022)

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

Krum and Krum $\circ$ Bucketing. Krum aggregation rule is defined as

$$\text{Krum}(x_1, \ldots, x_n) = \mathop{\arg\min}_{x_i \in \{x_1, \ldots, x_n\}} \sum_{j \in S_i} \|x_j - x_i\|^2,$$

where $S_i \subset \{x_1, \ldots, x_n\}$ is the subset of $n - B - 2$ closest vectors to x_i. By definition, $\text{Krum}(x_1, \ldots, x_n) \in \{x_1, \ldots, x_n\}$ and, thus $\|\text{Krum}(x_1, \ldots, x_n)\| \leqslant \max_{i\in[n]} \|x_i\|$, i.e., Assumption 1 holds with $F_{\mathcal{A}} = 1$. Since Krum $\circ$ Bucketing applies Krum aggregation to averages y_i over the buckets and $\|y_i\| \leqslant \frac{1}{s}\sum_{k=s(i-1)+1}^{\min\{si,n\}} \|x_{\pi(k)}\| \leqslant \max_{i\in[n]} \|x_i\|$, we have that $\|\text{Krum} \circ \text{Bucketing}(x_1, \ldots, x_n)\| \leqslant \max_{i\in[n]} \|x_i\|$.

Geometric median (GM) and GM $\circ$ Bucketing. Geometric median is defined as follows:

$$\text{GM}(x_1, \ldots, x_n) = \mathop{\arg\min}_{x \in \mathbb{R}^d} \sum_{i=1}^{n} \|x - x_i\|. \tag{14}$$

One can show that $\text{GM}(x_1, \ldots, x_n) \in \text{Conv}(x_1, \ldots, x_n) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d \mid x = \sum_{i=1}^n \alpha_i x_i \text{ for some } \alpha_1, \ldots, \alpha_n \geqslant 1 \text{ such that } \sum_{i=1}^n \alpha_i = 1\}$, i.e., geometric median belongs to the convex hull of the inputs. Indeed, let $\text{GM}(x_1, \ldots, x_n) = x = \hat{x} + \tilde{x}$, where $\hat{x}$ is the projection of x on $\text{Conv}(x_1, \ldots, x_n)$ and $\tilde{x} = x - \hat{x}$. Then, the optimality condition implies that $\langle \hat{x} - x, y - \hat{x} \rangle \geqslant 0$ for all $y \in \text{Conv}(x_1, \ldots, x_n)$. In particular, for all $i \in [n]$ we have $\langle \hat{x} - x, x_i - \hat{x} \rangle \geqslant 0$. Since

$$\begin{aligned}
\langle \hat{x} - x, x_i - \hat{x} \rangle &= \langle \tilde{x}, \hat{x} - x_i \rangle = \frac{1}{2}\|\tilde{x} + \hat{x} - x_i\|^2 - \frac{1}{2}\|\tilde{x}\|^2 - \frac{1}{2}\|\hat{x} - x_i\|^2 \\
&= \frac{1}{2}\|x - x_i\|^2 - \frac{1}{2}\|\tilde{x}\|^2 - \frac{1}{2}\|\hat{x} - x_i\|^2 \\
&\leqslant \frac{1}{2}\|x - x_i\|^2 - \frac{1}{2}\|\hat{x} - x_i\|^2,
\end{aligned}$$

we get that $\|x - x_i\| \geqslant \|\hat{x} - x_i\|$ for all $i \in [n]$ and the equality holds if and only if $\tilde{x} = 0$. Therefore, $\arg\min$ from equation 14 is achieved for x such that $x = \hat{x}$, meaning that $\text{GM}(x_1, \ldots, x_n) \in \text{Conv}(x_1, \ldots, x_n)$. Therefore, there exist some coefficients $\alpha_1, \ldots, \alpha_n \geqslant 0$ such that $\sum_{i=1}^n \alpha_i = 1$ and $\text{GM}(x_1, \ldots, x_n) = \sum_{i=1}^n \alpha_i x_i$, implying that

$$\|\text{GM}(x_1, \ldots, x_n)\| \leqslant \sum_{i=1}^{n} \alpha_i \|x_i\| \leqslant \max_{i\in[n]} \|x_i\|.$$

That is, GM satisfies Assumption 1 with $F_{\mathcal{A}} = 1$. Similarly to the case of Krum $\circ$ Bucketing, we also have $\|\text{GM} \circ \text{Bucketing}(x_1, \ldots, x_n)\| \leqslant \max_{i\in[n]} \|x_i\|$.

Coordinate-wise median (CM) and CM $\circ$ Bucketing. Coordinate-wise median (CM) is formally defined as

$$\text{CM}(x_1, \ldots, x_n) = \mathop{\arg\min}_{x \in \mathbb{R}^d} \sum_{i=1}^{n} \|x - x_i\|_1, \tag{15}$$

where $\|\cdot\|_1$ denotes ℓ_1-norm. This is equivalent to geometric median/median applied to vectors $x_1, \ldots, x_n$ component-wise. Therefore, from the above derivations for GM we have

$$\begin{aligned}
\|\text{CM}(x_1, \ldots, x_n)\|_\infty &\leqslant \max_{i\in[n]} \|x_i\|_\infty, \\
\|\text{CM} \circ \text{Bucketing}(x_1, \ldots, x_n)\|_\infty &\leqslant \max_{i\in[n]} \|x_i\|_\infty,
\end{aligned}$$

where $\|\cdot\|_\infty$ denotes ℓ_∞-norm. Therefore, due to the standard relations between ℓ_2- and ℓ_∞-norms, i.e., $\|a\|_\infty \leqslant \|a\| \leqslant \sqrt{d}\|a\|_\infty$ for any $a \in \mathbb{R}^d$, we have

$$\begin{aligned} \|\text{CM}(x_1, \ldots, x_n)\| &\leqslant \sqrt{d} \max_{i \in [n]} \|x_i\|, \\ \|\text{CM} \circ \text{Bucketing}(x_1, \ldots, x_n)\| &\leqslant \sqrt{d} \max_{i \in [n]} \|x_i\|, \end{aligned}$$

i.e., Assumption 1 is satisfied with $F_{\mathcal{A}} = \sqrt{d}$.

F Additional Convergence Results

In particular, the expression for A has 4 terms that depend on different smoothness parameters, client sampling (through $C, p_G, \mathcal{P}_{\mathcal{G}_C^k}$), ratio of Byzantine workers $\delta, \delta_{\max}$, overall number of clients, and communication compression parameter ω. Moreover, parameter p also implicitly depends on the client sampling and communication compression: to make the expected number of clients participating in the communication round equal to $\mathcal{O}(C)$, to make the expected number of stochastic oracle calls equal to $\mathcal{O}(b)$, and to make the expected number of transmitted components for each worker taking part in the communication round equal $\mathcal{O}(\zeta_Q)$, parameter p should be chosen as $p = \min\{C/n, b/m, \zeta_Q/d\}$, where the latter term in the minimum often equals to $\Theta(1/(\omega+1))$ (Gorbunov et al., 2021).

As we explain in the appendix (see the discussion after Theorem G.1), when $\omega = 0$ (no compression), the bound for A is proportional to $1/p^2$ in the case of Byz-VR-MARINA-PP even when $C = \widehat{C} = n$ (full participation) and $\delta = 0$ (it is known apriori that there are no Byzantine workers). In contrast, a similar quantity in the result for Byz-VR-MARINA (Gorbunov et al., 2023) is proportional to $1/p$. In this special case, we do not recover the result for Byz-VR-MARINA.

Such a complexity deterioration can be explained as follows: the presence of clipping introduces additional technical difficulties in the analysis, resulting in a reduced step size compared to Byz-VR-MARINA, even when $C = n$. To achieve a more favorable convergence rate, particularly in scenarios of complete participation, we also establish the results under the assumptions detailed in Assumption 2 below.

Theorem F.1. *Let Assumptions 1, 2, 3, 4, 5, 6, 7 hold and* $\lambda_{k+1} = D_Q \max_{i,j} L_{i,j} \left\|x^{k+1} - x^k\right\|$. *Assume that* $0 < \gamma \leqslant 1/(L+\sqrt{A})$, *where constant A equals*

$$\begin{aligned}
A &= A_1 L^2 + A_2 \max_{i,j} L_{i,j}^2 + A_3 L_{\pm}^2 + \frac{\mathcal{L}_{\pm}^2}{b}, \\
A_1 &= \frac{2p_G \mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{pC(1-\delta_{\max})}\omega + \frac{8}{p^2}(1-p_G) \\
&\quad + \frac{16}{p^2} p_G \mathcal{P}_{\mathcal{G}_C^k} c\delta_{\max}\omega, \\
A_2 &= \frac{8}{p^2}(1-p_G) F_{\mathcal{A}}^2 D_Q^2, \\
A_3 &= \frac{2p_G \mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{pC(1-\delta_{\max})}(\omega+1) \\
&\quad + \frac{16}{p^2} p_G \mathcal{P}_{\mathcal{G}_C^k} c\delta_{\max}(\omega+1), \\
A_4 &= \frac{2p_G \mathcal{P}_{\mathcal{G}_C^k}(1+\omega)c\delta_{\max}}{p} \\
&\quad + \frac{16}{p^2} p_G \mathcal{P}_{\mathcal{G}_C^k}(1+\omega)\frac{(1-\delta)n}{C(1-\delta_{\max})} \\
\widehat{C} &= 2\frac{\delta_{\max}\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}}{(1-\delta_{\max})} B(6c\delta_{\max}+1)
\end{aligned}$$

Then for all $K \geqslant 0$ the iterates produced by Byz-VR-MARINA-PP *(Algorithm 1) satisfy*

$$\mathbb{E}\left[\left\|\nabla f\left(\widehat{x}^K\right)\right\|^2\right] \leqslant \frac{2\Phi^0}{\gamma(K+1)} + \frac{2\widehat{C}\zeta^2}{p}, \tag{16}$$

where $\widehat{x}^K$ is choosen uniformly at random from $x^0, x^1, \ldots, x^K$, and $\Phi^0 = f\left(x^0\right) - f^* + \frac{\gamma}{p}\left\|g^0 - \nabla f\left(x^0\right)\right\|^2$. *If, in addition, Assumption 8 holds and* $0 < \gamma \leqslant 1/(L+\sqrt{2A})$, *then for all $K \geqslant 0$ the iterates produced by* Byz-VR-MARINA-PP *(Algorithm 1) satisfy with* $\rho = \min\left\{\gamma\mu, \frac{p}{4}\right\}$

$$\mathbb{E}\left[f\left(x^K\right) - f\left(x^*\right)\right] \leqslant (1-\rho)^K \Phi^0 + \frac{2\widehat{C}\zeta^2\gamma}{p\rho}, \tag{17}$$

where $\Phi^0 = f\left(x^0\right) - f^* + \frac{2\gamma}{p}\left\|g^0 - \nabla f\left(x^0\right)\right\|^2$.

With Assumptions 2 and 7, vectors $\{\mathcal{Q}(\widehat{\Delta}_i(x^{k+1}, x^k))\}_{i\in\mathcal{G}_C^k}$ can be upper bounded by $D_Q \max_{i,j} L_{i,j} \left\|x^{k+1} - x^k\right\|$. Using this fact, one can take the clipping level sufficiently large such that it is turned off for the regular workers. This allows us to simplify the proof and remove $1/p$ factor in front of the terms not proportional to $\delta_{\max}$ or to $1 - p_G$ in the expression for A that can make the stepsize larger. However, the formula for the constant A also contains the term $\frac{8}{p^2}(1 - p_G)D_Q^2 \max_{i,j} L_{i,j}^2$ that is larger than the corresponding term from Theorem 4.1. When D_Q is large or when $\max_{i,j} L_{i,j}$ is much greater than $\max_i L_i$, the stepsize from Theorem F.1 can be even smaller than the one from Theorem 4.1. Therefore, the rates of convergence cannot be compared directly. We also highlight that the clipping level from Theorem F.1 is in general larger than the clipping level from Theorem 4.1 and, thus, it is expected that with participation Theorem F.1 gives better results than Theorem 4.1: the bias introduced due to the clipping becomes smaller with the increase of the clipping level. However, in the partial participation regime, the price for this is a potential decrease of the stepsize to compensate for the increased harm from Byzantine workers in the situations when they form a majority.

G General Analysis

Lemma G.1. *Let X be a random vector in $\mathbb{R}^d$ and $\widetilde{X} = \text{clip}_\lambda(X)$. Assume that $\mathbb{E}[X] = x \in \mathbb{R}^d$ and $\|x\| \leqslant \lambda/2$, then*

$$\mathbb{E}\left[\|\widetilde{X} - x\|^2\right] \leqslant 10\mathbb{E}\,\|X - x\|^2 .$$

Proof. The proof follows a similar procedure to that presented in Lemma F.5 from (Gorbunov et al., 2020). To commence the proof, we introduce two indicator random variables:

$$\chi = \mathbb{I}_{\{X:\|X\|>\lambda\}} = \begin{cases} 1, & \text{if } \|X\| > \lambda, \\ 0, & \text{otherwise} \end{cases}, \eta = \mathbb{I}_{\{X:\|X-x\|>\frac{\lambda}{2}\}} = \begin{cases} 1, & \text{if } \|X - x\| > \frac{\lambda}{2} \\ 0, & \text{otherwise} \end{cases}.$$

Moreover, since $\|X\| \leqslant \|x\| + \|X - x\| \overset{\|x\|\leqslant\lambda/2}{\leqslant} \frac{\lambda}{2} + \|X - x\|$, we have $\chi \leqslant \eta$. Using that we get

$$\widetilde{X} = \min\left\{1, \frac{\lambda}{\|X\|}\right\} X = \chi\frac{\lambda}{\|X\|}X + (1-\chi)X.$$

By Markov's inequality,

$$\mathbb{E}[\eta] = \mathbb{P}\left\{\|X - x\| > \frac{\lambda}{2}\right\} = \mathbb{P}\left\{\|X - x\|^2 > \frac{\lambda^2}{4}\right\} \leqslant \frac{4}{\lambda^2}\mathbb{E}\left[\|X - x\|^2\right]. \tag{18}$$

Using $\|\widetilde{X} - x\| \leqslant \|\widetilde{X}\| + \|x\| \leqslant \lambda + \frac{\lambda}{2} = \frac{3\lambda}{2}$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\widetilde{X} - x\|^2\right] &= \mathbb{E}\left[\|\widetilde{X} - x\|^2\chi + \|\widetilde{X} - x\|^2(1-\chi)\right] \\
&= \mathbb{E}\left[\chi\left\|\frac{\lambda}{\|X\|}X - x\right\|^2 + \|X - x\|^2(1-\chi)\right] \\
&\leqslant \mathbb{E}\left[\chi\left(\left\|\frac{\lambda}{\|X\|}X\right\| + \|x\|\right)^2 + \|X - x\|^2(1-\chi)\right] \\
&\overset{\|x\|\leqslant\frac{\lambda}{2}}{\leqslant} \left(\mathbb{E}\left[\chi\left(\frac{3\lambda}{2}\right)^2 + \|X - x\|^2\right]\right),
\end{aligned}$$

where in the last inequality we applied $1 - \chi \leqslant 1$. Using (18) and $\chi \leqslant \eta$ we get

$$\begin{aligned}
\mathbb{E}\left[\|\widetilde{X} - x\|^2\right] &\leqslant \frac{9\lambda^2}{4}\left(\frac{2}{\lambda}\right)^2\mathbb{E}\left[\|X - x\|^2\right] + \mathbb{E}\left[\|X - x\|^2\right] \\
&\leqslant 10\mathbb{E}\left[\|X - x\|^2\right].
\end{aligned}$$

□

Lemma G.2 (Lemma 2 from Li et al. (2021))**.** *Assume that function f is L-smooth (Assumption 4) and $x^{k+1} = x^k - \gamma g^k$. Then*

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

Next, instead of Assumption 3, we consider a more generalized one.

Assumption 9 ((B, ζ^2)-heterogeneity)**.** *We assume that good clients have $\left(B, \zeta^2\right)$-heterogeneous local loss functions for some $B \geqslant 0, \zeta \geqslant 0$, i.e.,*

$$\frac{1}{G}\sum_{i\in\mathcal{G}}\|\nabla f_i(x) - \nabla f(x)\|^2 \leqslant B\|\nabla f(x)\|^2 + \zeta^2 \quad \forall x \in \mathbb{R}^d$$

When $B = 0$, the above assumption recovers Assumption 3. However, it also covers some situations when the model is over-parameterized (Vaswani et al., 2019) and can hold with smaller values of ζ^2. This assumption is also used in (Karimireddy et al., 2022; Gorbunov et al., 2023).

Lemma G.3. *Let Assumptions 4, 5, 6 hold and the Compression Operator satisfy Definition 2.2. Let us define "ideal" estimator:*

$$\overline{g}^{k+1} = \begin{cases} \frac{1}{G_c^k}\sum\limits_{i\in\mathcal{G}_c^k}\nabla f_i(x^{k+1}), & c_n = 1, & [1] \\ g^k + \nabla f\left(x^{k+1}\right) - \nabla f\left(x^k\right), & c_n = 0 \text{ and } G_c^k < (1-\delta_{\max})C, & [2] \\ g^k + \frac{1}{G_c^k}\sum\limits_{i\in\mathcal{G}_C^k}\operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right), & c_n = 0 \text{ and } G_c^k \geqslant (1-\delta_{\max})C. & [3] \end{cases}$$

Then for all $k \geqslant 0$ the iterates produced by Byz-VR-MARINA-PP *(Algorithm 1) satisfy*

$$\begin{aligned} A_1 &= \mathbb{E}\left[\left\|\overline{g}^{k+1} - \nabla f\left(x^{k+1}\right)\right\|^2\right] \\ &\leqslant (1-p)\left(1+\frac{p}{4}\right)\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] + p\frac{\delta_{\max}\cdot\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}}{(1-\delta_{\max})}\mathbb{E}\left[B\|\nabla f(x)\|^2 + \zeta^2\right] \\ &+ (1-p)p_G\left(1+\frac{4}{p}\right)\frac{2\cdot\mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}\left(10\omega L^2 + (10\omega+1)L_\pm^2 + \frac{10(\omega+1)\mathcal{L}_\pm^2}{b}\right)\mathbb{E}\left[\|x^{k+1}-x^k\|^2\right], \end{aligned}$$

where $p_G = \operatorname{Prob}\left\{G_C^k \geqslant (1-\delta_{\max})C\right\}$ and $\mathcal{P}_{\mathcal{G}_C^k} = \operatorname{Prob}\left\{i\in\mathcal{G}_C^k \mid G_C^k \geqslant (1-\delta_{\max})\,C\right\}$.

Proof. Let us examine the expected value of the squared difference between ideal estimator and full gradient:

$$\begin{aligned} A_1 &= \mathbb{E}\left[\left\|\overline{g}^{k+1} - \nabla f\left(x^{k+1}\right)\right\|^2\right] \\ &= \mathbb{E}\left[\mathbb{E}_k\left[\left\|\overline{g}^{k+1} - \nabla f\left(x^{k+1}\right)\right\|^2\right]\right] \\ &= (1-p)\,p_G\mathbb{E}\left[\mathbb{E}_k\left[\left\|g^k + \frac{1}{G_c^k}\sum_{i\in\mathcal{G}_C^k}\operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right) - \nabla f\left(x^{k+1}\right)\right\|^2\right] \mid [3]\right] \\ &+ (1-p)(1-p_G)\mathbb{E}\left[\mathbb{E}_k\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \mid [2]\right] + p\mathbb{E}\left[\left\|\frac{1}{G_c^k}\sum_{i\in\mathcal{G}_c^k}\nabla f_i(x^{k+1}) - \nabla f(x^{k+1})\right\|^2\right]. \end{aligned}$$

Using (7) and $\nabla f\left(x^k\right) - \nabla f\left(x^k\right) = 0$ we obtain

$$\begin{aligned} B_1 &= \mathbb{E}\left[\mathbb{E}_k\left[\left\|g^k + \frac{1}{G_c^k}\sum_{i\in\mathcal{G}_C^k}\operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right) - \nabla f\left(x^{k+1}\right)\right\|^2\right] \mid [3]\right] \\ &= \mathbb{E}\left[\mathbb{E}_k\left[\left\|g^k + \frac{1}{G_c^k}\sum_{i\in\mathcal{G}_C^k}\operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right) - \nabla f\left(x^{k+1}\right) + \nabla f\left(x^k\right) - \nabla f\left(x^k\right)\right\|^2\right] \mid [3]\right] \\ &\overset{(7)}{\leqslant} \left(1+\frac{p}{4}\right)\mathbb{E}\left[\left\|g^k - \nabla f\left(x^k\right)\right\|^2\right] \\ &+ \left(1+\frac{4}{p}\right)\mathbb{E}\left[\mathbb{E}_k\left[\left\|\frac{1}{G_c^k}\sum_{i\in\mathcal{G}_C^k}\operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right) - \left(\nabla f(x^{k+1}) - \nabla f(x^k)\right)\right\|^2\right] \mid [3]\right] \\ &= \left(1+\frac{p}{4}\right)\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \\ &+ \left(1+\frac{4}{p}\right)\mathbb{E}\left[\mathbb{E}_k\left[\left\|\frac{1}{G_c^k}\sum_{i\in\mathcal{G}_C^k}\operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right) - \Delta\left(x^{k+1},x^k\right)\right\|^2\right] \mid [3]\right]. \end{aligned}$$

Let us consider last part of the inequality:

$$B_1' = \mathbb{E}\left[\mathbb{E}_k\left[\left\|\frac{1}{G_c^k}\sum_{i\in\mathcal{G}_C^k}\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2\right]\mid[3]\right]$$
$$= \mathbb{E}\left[\mathbb{E}_{S_k}\left[\mathbb{E}_k\left[\left\|\frac{1}{G_c^k}\sum_{i\in\mathcal{G}_C^k}\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2\right]\mid[3]\right]\right].$$

Note that $G_C^k \geqslant (1-\delta_{\max})C$ in this case:

$$B_1' \leqslant \frac{1}{C(1-\delta_{\max})}\mathbb{E}\left[\mathbb{E}_{S_k}\left[\sum_{i\in\mathcal{G}_C^k}\mathbb{E}_k\left[\left\|\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2\right]\mid[3]\right]\right]$$
$$\leqslant \frac{1}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\mathbb{E}_{S_k}\left[\mathcal{I}_{\mathcal{G}_C^k}\right]\mathbb{E}_k\left[\left\|\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2\right]\mid[3]\right]$$
$$= \frac{1}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\mathcal{P}_{\mathcal{G}_C^k}\cdot\mathbb{E}_k\left[\left\|\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2\right]\mid[3]\right], \quad (19)$$

where $\mathcal{I}_{\mathcal{G}_C^k}$ is an indicator function for the event $\left\{i\in\mathcal{G}_C^k \mid G_C^k \geqslant (1-\delta_{\max})\,C\right\}$ and $\mathcal{P}_{\mathcal{G}_C^k} = \text{Prob}\left\{i\in\mathcal{G}_C^k \mid G_C^k \geqslant (1-\delta_{\max})\,C\right\}$ is probability of such event. Note that $\mathbb{E}_{S_k}\left[\mathcal{I}_{\mathcal{G}_C^k}\right] = \mathcal{P}_{\mathcal{G}_C^k}$. In case of uniform sampling of clients we have

$$\forall i\in\mathcal{G}\quad \mathcal{P}_{\mathcal{G}_C^k} = \text{Prob}\left\{i\in\mathcal{G}_C^k \mid G_C^k \geqslant (1-\delta_{\max})\,C\right\}$$
$$= \frac{C}{np_G}\cdot\sum_{(1-\delta_{\max})C\leqslant t\leqslant C}\left(\binom{G}{t}\binom{n-G}{C-t}\left(\binom{n}{C}\right)^{-1}\right),$$
$$p_G = \sum_{(1-\delta_{\max})C\leqslant t\leqslant C}\left(\binom{G-1}{t-1}\binom{n-G}{C-t}\left(\binom{n-1}{C-1}\right)^{-1}\right)$$

Now we can continue with inequalities:

$$B_1' \leqslant \frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\mathbb{E}_k\left[\left\|\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2\right]\mid[3]\right]$$
$$\leqslant \frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\mathbb{E}_k\left[\mathbb{E}_Q\left[\left\|\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2\right]\right]\mid[3]\right]$$
$$\overset{(7)}{\leqslant} \frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}2\mathbb{E}_k\left[\mathbb{E}_Q\left[\left\|\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2\right]\right]\mid[3]\right]$$
$$+ \frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}2\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2\right]\mid[3]\right].$$

Using Lemma G.1 we have

$$B_1' \overset{\text{Lemma G.1}}{\leqslant} \frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}} 20\mathbb{E}_k\left[\mathbb{E}_Q\left[\left\|\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2\right]\right] \mid [3]\right]$$
$$+\frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}} 2\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2\right] \mid [3]\right]$$
$$\leqslant \frac{20\cdot\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}} \mathbb{E}_k\left[\mathbb{E}_Q\left[\left\|\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2\right]\right] \mid [3]\right]$$
$$+\frac{2\cdot\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}} \mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2\right] \mid [3]\right]$$
$$\leqslant \frac{20\cdot\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}} \mathbb{E}_k\left[\mathbb{E}_Q\left[\left\|\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right\|^2\right]\right]-\sum_{i\in\mathcal{G}}\left\|\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]$$
$$+\frac{2\cdot\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}} \mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2\right] \mid [3]\right].$$

Applying Definition 2.2 of Unbiased Compressor we have

$$B_1' \leqslant \frac{20\cdot\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}(1+\omega)\mathbb{E}_k\left\|\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right\|^2-\sum_{i\in\mathcal{G}}\left\|\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]$$
$$+\frac{2\cdot\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]$$
$$\leqslant \frac{20\cdot\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}(1+\omega)\mathbb{E}_k\left\|\widehat{\Delta}_i\left(x^{k+1},x^k\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2\right]$$
$$+\frac{20\cdot\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}(1+\omega)\mathbb{E}_k\left\|\Delta_i\left(x^{k+1},x^k\right)\right\|^2-\sum_{i\in\mathcal{G}}\mathbb{E}_k\left\|\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]$$
$$+\frac{2\cdot\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right].$$

Now we combine terms and have

$$B_1' \leqslant \frac{20\cdot\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}(1+\omega)\mathbb{E}\left[\sum_{i\in\mathcal{G}}\mathbb{E}_k\left[\left\|\widehat{\Delta}_i\left(x^{k+1},x^k\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2\right] \mid [3]\right]$$
$$+\frac{20\cdot\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\omega\mathbb{E}\left[\sum_{i\in\mathcal{G}}\left\|\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]$$
$$+\frac{2\cdot\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]$$
$$=\frac{20\cdot\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}(1+\omega)\mathbb{E}\left[\sum_{i\in\mathcal{G}}\mathbb{E}_k\left[\left\|\widehat{\Delta}_i\left(x^{k+1},x^k\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2\right] \mid [3]\right]$$
$$+\frac{20\cdot\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\omega\mathbb{E}\left[\sum_{i\in\mathcal{G}}\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2+\left\|\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]$$
$$+\frac{2\cdot\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right].$$

Rearranging terms leads to

$$
\begin{aligned}
B_1' \leqslant & \frac{20 \cdot \mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}(1+\omega)\mathbb{E}\left[\sum_{i\in\mathcal{G}}\mathbb{E}_k\left[\left\|\widehat{\Delta}_i\left(x^{k+1},x^k\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2\right] \mid [3]\right] \\
&+ \frac{2 \cdot \mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}(10\omega+1)\mathbb{E}\left[\sum_{i\in\mathcal{G}}\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right] \\
&+ \frac{20 \cdot \mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\omega\mathbb{E}\left[\sum_{i\in\mathcal{G}}\left\|\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right].
\end{aligned}
$$

Now we apply Assumptions 4, 5, 6:

$$
\begin{aligned}
B_1' \leqslant & \frac{20 \cdot \mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}(1+\omega)\mathbb{E}\left[G\frac{\mathcal{L}_\pm^2}{b}\|x^{k+1}-x^k\|^2\right] \\
&+ \frac{2 \cdot \mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}(10\omega+1)\mathbb{E}\left[GL_\pm^2\|x^{k+1}-x^k\|^2\right] \\
&+ \frac{20 \cdot \mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\omega\mathbb{E}\left[GL^2\left\|x^{k+1}-x^k\right\|^2\right].
\end{aligned}
$$

Finally, we have

$$
B_1' \leqslant \frac{2 \cdot \mathcal{P}_{\mathcal{G}_C^k}\cdot G}{C(1-\delta_{\max})}\left(10\omega L^2 + (10\omega+1)L_\pm^2 + \frac{10(\omega+1)\mathcal{L}_\pm^2}{b}\right)\mathbb{E}\left[\|x^{k+1}-x^k\|^2\right].
$$

Let us plug obtained results:

$$
\begin{aligned}
B_1 \leqslant & \left(1+\frac{p}{4}\right)\mathbb{E}\left[\left\|g^k-\nabla f(x^k)\right\|^2\right] \\
&+ \left(1+\frac{4}{p}\right)\frac{2 \cdot \mathcal{P}_{\mathcal{G}_C^k}\cdot G}{C(1-\delta_{\max})}\left(10\omega L^2 + (10\omega+1)L_\pm^2 + \frac{10(\omega+1)\mathcal{L}_\pm^2}{b}\right)\mathbb{E}\left[\|x^{k+1}-x^k\|^2\right].
\end{aligned}
$$

Let us consider the term $\mathbb{E}\left[\left\|\frac{1}{G_c^k}\sum\limits_{i\in\mathcal{G}_c^k}\nabla f_i(x^{k+1})-\nabla f(x^{k+1})\right\|^2\right]$:

$$
\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{G_c^k}\sum_{i\in\mathcal{G}_c^k}\nabla f_i(x^{k+1})-\nabla f(x^{k+1})\right\|^2\right] &\leqslant \mathbb{E}\left[\frac{1}{G_c^k}\sum_{i\in\mathcal{G}_c^k}\left\|\nabla f_i(x^{k+1})-\nabla f(x^{k+1})\right\|^2\right] \\
&\leqslant \frac{1}{(1-\delta_{\max})\widehat{C}}\mathbb{E}\left[\sum_{i\in\mathcal{G}_c^k}\left\|\nabla f_i(x^{k+1})-\nabla f(x^{k+1})\right\|^2\right] \\
&= \frac{1}{(1-\delta_{\max})\widehat{C}}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\mathcal{I}_{\mathcal{G}_{\widehat{C}}^k}\left\|\nabla f_i(x^{k+1})-\nabla f(x^{k+1})\right\|^2\right]
\end{aligned}
$$

Using definition of $\mathcal{P}_{\mathcal{G}_C^k}$ we get

$$
\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{G_c^k}\sum_{i\in\mathcal{G}_c^k}\nabla f_i(x^{k+1})-\nabla f(x^{k+1})\right\|^2\right] &\leqslant \frac{\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}}{(1-\delta_{\max})\widehat{C}}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\left\|\nabla f_i(x^{k+1})-\nabla f(x^{k+1})\right\|^2\right] \\
&\leqslant \frac{G\cdot\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}}{(1-\delta_{\max})\widehat{C}G}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\left\|\nabla f_i(x^{k+1})-\nabla f(x^{k+1})\right\|^2\right]
\end{aligned}
$$

Using Assumption 9 we get

$$\mathbb{E}\left[\left\|\frac{1}{G_c^k}\sum_{i\in\mathcal{G}_c^k}\nabla f_i(x^{k+1})-\nabla f(x^{k+1})\right\|^2\right] \leqslant \frac{G\cdot\mathcal{P}_{\mathcal{G}^k_{\widehat{C}}}}{(1-\delta_{\max})\widehat{C}}\mathbb{E}\left[B\|\nabla f(x)\|^2+\zeta^2\right]$$
$$= \frac{\delta n\cdot\mathcal{P}_{\mathcal{G}^k_{\widehat{C}}}}{(1-\delta_{\max})\frac{\delta n}{\delta_{\max}}}\mathbb{E}\left[B\|\nabla f(x)\|^2+\zeta^2\right]$$
$$= \frac{\delta_{\max}\cdot\mathcal{P}_{\mathcal{G}^k_{\widehat{C}}}}{(1-\delta_{\max})}\mathbb{E}\left[B\|\nabla f(x)\|^2+\zeta^2\right]$$

Also we have

$$\begin{aligned}
A_1 &= \mathbb{E}\left[\left\|\overline{g}^{k+1}-\nabla f(x^{k+1})\right\|^2\right]\\
&\leqslant (1-p)p_G B_1 + (1-p)(1-p_G)\mathbb{E}\left[\left\|g^k-\nabla f(x^k)\right\|^2\right] + p\frac{\delta_{\max}\cdot\mathcal{P}_{\mathcal{G}^k_{\widehat{C}}}}{(1-\delta_{\max})}\mathbb{E}\left[B\|\nabla f(x)\|^2+\zeta^2\right]\\
&\leqslant (1-p)p_G\left(1+\frac{p}{4}\right)\mathbb{E}\left[\left\|g^k-\nabla f(x^k)\right\|^2\right]\\
&+ (1-p)p_G\left(1+\frac{4}{p}\right)\frac{2\cdot\mathcal{P}_{\mathcal{G}^k_{C}}\cdot G}{C(1-\delta_{\max})}\left(10\omega L^2+(10\omega+1)L_{\pm}^2+\frac{10(\omega+1)\mathcal{L}_{\pm}^2}{b}\right)\mathbb{E}\left[\|x^{k+1}-x^k\|^2\right]\\
&+ (1-p)(1-p_G)\mathbb{E}\left[\left\|g^k-\nabla f(x^k)\right\|^2\right] + p\frac{\delta_{\max}\cdot\mathcal{P}_{\mathcal{G}^k_{\widehat{C}}}}{(1-\delta_{\max})}\mathbb{E}\left[B\|\nabla f(x)\|^2+\zeta^2\right].
\end{aligned}$$

To simplify the bound we use $\left(1+\frac{p}{4}>1\right)$ and obtain

$$\begin{aligned}
A_1 &\leqslant (1-p)p_G\left(1+\frac{p}{4}\right)\mathbb{E}\left[\left\|g^k-\nabla f(x^k)\right\|^2\right] + p\frac{\delta_{\max}\cdot\mathcal{P}_{\mathcal{G}^k_{\widehat{C}}}}{(1-\delta_{\max})}\mathbb{E}\left[B\|\nabla f(x)\|^2+\zeta^2\right]\\
&+ (1-p)p_G\left(1+\frac{4}{p}\right)\frac{2\cdot\mathcal{P}_{\mathcal{G}^k_{C}}\cdot G}{C(1-\delta_{\max})}\left(10\omega L^2+(10\omega+1)L_{\pm}^2+\frac{10(\omega+1)\mathcal{L}_{\pm}^2}{b}\right)\mathbb{E}\left[\|x^{k+1}-x^k\|^2\right]\\
&+ (1-p)(1-p_G)\mathbb{E}\left[\left\|g^k-\nabla f(x^k)\right\|^2\right]\\
&\leqslant (1-p)p_G\left(1+\frac{p}{4}\right)\mathbb{E}\left[\left\|g^k-\nabla f(x^k)\right\|^2\right]\\
&+ (1-p)p_G\left(1+\frac{4}{p}\right)\frac{2\cdot\mathcal{P}_{\mathcal{G}^k_{C}}\cdot G}{C(1-\delta_{\max})}\left(10\omega L^2+(10\omega+1)L_{\pm}^2+\frac{10(\omega+1)\mathcal{L}_{\pm}^2}{b}\right)\mathbb{E}\left[\|x^{k+1}-x^k\|^2\right]\\
&+ (1-p)(1-p_G)\left(1+\frac{p}{4}\right)\mathbb{E}\left[\left\|g^k-\nabla f(x^k)\right\|^2\right] + p\frac{\delta_{\max}\cdot\mathcal{P}_{\mathcal{G}^k_{\widehat{C}}}}{(1-\delta_{\max})}\mathbb{E}\left[B\|\nabla f(x)\|^2+\zeta^2\right]\\
&\leqslant (1-p)\left(1+\frac{p}{4}\right)\mathbb{E}\left[\left\|g^k-\nabla f(x^k)\right\|^2\right] + p\frac{\delta_{\max}\cdot\mathcal{P}_{\mathcal{G}^k_{\widehat{C}}}}{(1-\delta_{\max})}\mathbb{E}\left[B\|\nabla f(x)\|^2+\zeta^2\right]\\
&+ (1-p)p_G\left(1+\frac{4}{p}\right)\frac{2\cdot\mathcal{P}_{\mathcal{G}^k_{C}}(1-\delta)n}{C(1-\delta_{\max})}\left(10\omega L^2+(10\omega+1)L_{\pm}^2+\frac{10(\omega+1)\mathcal{L}_{\pm}^2}{b}\right)\mathbb{E}\left[\|x^{k+1}-x^k\|^2\right].
\end{aligned}$$

□

Lemma G.4. *Let us define "ideal" estimator:*

$$\overline{g}^{k+1} = \begin{cases} \frac{1}{G_c^k}\sum\limits_{i\in\mathcal{G}_c^k}\nabla f_i(x^{k+1}), & c_n = 1, & [1]\\ g^k+\nabla f\left(x^{k+1}\right)-\nabla f\left(x^k\right), & c_n = 0 \text{ and } G_c^k < (1-\delta_{\max})C, & [2]\\ g^k+\frac{1}{G_c^k}\sum\limits_{i\in\mathcal{G}_C^k}\operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right), & c_n = 0 \text{ and } G_c^k \geqslant (1-\delta_{\max})C. & [3]\end{cases}$$

Also let us introduce the notation

$$\texttt{ARAgg}_Q^{k+1} = \texttt{ARAgg}\left(\operatorname{clip}_{\lambda_{k+1}}\left(\mathcal{Q}\left(\widehat{\Delta}_1(x^{k+1},x^k)\right)\right),\ldots,\operatorname{clip}_{\lambda_{k+1}}\left(\mathcal{Q}\left(\widehat{\Delta}_C(x^{k+1},x^k)\right)\right)\right).$$

Then for all $k \geqslant 0$ the iterates produced by Byz-VR-MARINA-PP *(Algorithm 1) satisfy*

$$
\begin{aligned}
A_2 &= \mathbb{E}\left[\left\|g^{k+1} - \overline{g}^{k+1}\right\|^2\right] \\
&\leqslant p\mathbb{E}\left[\mathbb{E}_k\left[\left\|\texttt{ARAgg}\left(\nabla f_1(x^{k+1}), \ldots, \nabla f_{\widehat{C}}(x^{k+1})\right) - \nabla f(x^{k+1})\right\|^2\right] \mid [1]\right] \\
&\quad + (1-p)p_G\mathbb{E}\left[\mathbb{E}_k\left[\left\|\frac{1}{G_c^k}\sum_{i\in\mathcal{G}_C^k}\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1}, x^k\right)\right)\right) - \texttt{ARAgg}_Q^{k+1}\right\|^2 \mid [3]\right]\right] \\
&\quad + (1-p)(1-p_G)\mathbb{E}\left[\mathbb{E}_k\left[\left\|\nabla f(x^{k+1}) - \nabla f(x^k) - \texttt{ARAgg}_Q^{k+1}\right\|^2 \mid [2]\right]\right],
\end{aligned}
$$

where $p_G = \text{Prob}\left\{G_C^k \geqslant (1-\delta_{\max})C\right\}$.

Proof. Using conditional expectations we have

$$
\begin{aligned}
A_2 &= \mathbb{E}\left[\mathbb{E}_k\left[\left\|g^{k+1} - \overline{g}^{k+1}\right\|^2\right]\right] \\
&= p\mathbb{E}\left[\mathbb{E}_k\left[\left\|\texttt{ARAgg}\left(\nabla f_1(x^{k+1}), \ldots, \nabla f_{\widehat{C}}(x^{k+1})\right) - \nabla f(x^{k+1})\right\|^2\right] \mid [1]\right] \\
&+ (1-p)p_G\mathbb{E}\left[\mathbb{E}_k\left[\left\|g^k + \frac{1}{G_c^k}\sum_{i\in\mathcal{G}_C^k}\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1}, x^k\right)\right)\right) - \left(g^k + \texttt{ARAgg}_Q^{k+1}\right)\right\|^2\right] \mid [3]\right] \\
&\quad + (1-p)(1-p_G)\mathbb{E}\left[\mathbb{E}_k\left[\left\|g^k + \nabla f(x^{k+1}) - \nabla f(x^k) - \left(g^k + \texttt{ARAgg}_Q^{k+1}\right)\right\|^2\right] \mid [2]\right].
\end{aligned}
$$

After simplification we get the following bound:

$$
\begin{aligned}
A_2 &\leqslant p\mathbb{E}\left[\mathbb{E}_k\left[\left\|\texttt{ARAgg}\left(\nabla f_1(x^{k+1}), \ldots, \nabla f_{\widehat{C}}(x^{k+1})\right) - \nabla f(x^{k+1})\right\|^2\right] \mid [1]\right] \\
&\quad + (1-p)p_G\mathbb{E}\left[\mathbb{E}_k\left[\left\|\frac{1}{G_c^k}\sum_{i\in\mathcal{G}_C^k}\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1}, x^k\right)\right)\right) - \texttt{ARAgg}_Q^{k+1}\right\|^2 \mid [3]\right]\right] \\
&\quad + (1-p)(1-p_G)\mathbb{E}\left[\mathbb{E}_k\left[\left\|\nabla f(x^{k+1}) - \nabla f(x^k) - \texttt{ARAgg}_Q^{k+1}\right\|^2 \mid [2]\right]\right].
\end{aligned}
$$

□

Lemma G.5. *Let Assumptions 4 and 9 hold and Aggregation Operator (*ARAgg*) satisfy Definition 2.1. Then for all $k \geqslant 0$ the iterates produced by* Byz-VR-MARINA-PP *(Algorithm 1) satisfy*

$$
\begin{aligned}
T_1 &= \mathbb{E}\left[\mathbb{E}_k\left[\left\|\texttt{ARAgg}\left(\nabla f_1(x^{k+1}), \ldots, \nabla f_{\widehat{C}}(x^{k+1})\right) - \nabla f(x^{k+1})\right\|^2\right] \mid [1]\right] \\
&\leqslant 8\frac{\delta n \mathcal{P}_{\mathcal{G}_{\widehat{C}}^k} c\delta_{\max}}{(1-\delta_{\max})\frac{\delta n}{\delta_{\max}}} B\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] + 8\frac{\delta n \mathcal{P}_{\mathcal{G}_{\widehat{C}}^k} c\delta_{\max}}{(1-\delta_{\max})\frac{\delta n}{\delta_{\max}}} BL^2\mathbb{E}\left[\left\|x^{k+1} - x^k\right\|^2\right] + 4\frac{\delta n \mathcal{P}_{\mathcal{G}_{\widehat{C}}^k} c\delta_{\max}}{(1-\delta_{\max})\frac{\delta n}{\delta_{\max}}}\zeta^2.
\end{aligned}
$$

Proof. Using Definition of aggregation operator we have

$$
\begin{aligned}
T_1 &= \mathbb{E}\left[\mathbb{E}_k\left[\left\|\texttt{ARAgg}\left(\nabla f_1(x^{k+1}),\ldots,\nabla f_{\widehat{C}}(x^{k+1})\right)-\nabla f(x^{k+1})\right\|^2 \mid [1]\right]\right]\\
&\overset{\text{(Def. 2.1)}}{\leqslant} \mathbb{E}\left[\frac{c\delta_{\max}}{G_{\widehat{C}}^k(G_{\widehat{C}}^k-1)}\sum_{\substack{i,l\in\mathcal{G}_{\widehat{C}}^k\\ i\neq l}}\mathbb{E}_k\left[\left\|\nabla f_i\left(x^{k+1}\right)-\nabla f_l\left(x^{k+1}\right)\right\|^2 \mid [1]\right]\right]\\
&\overset{(7)}{\leqslant} \mathbb{E}\left[\frac{c\delta_{\max}}{G_{\widehat{C}}^k(G_{\widehat{C}}^k-1)}\sum_{\substack{i,l\in\mathcal{G}_{\widehat{C}}^k\\ i\neq l}}\mathbb{E}\left[2\left\|\nabla f_i\left(x^{k+1}\right)-\nabla f\left(x^{k+1}\right)\right\|^2+2\left\|\nabla f_l\left(x^{k+1}\right)-\nabla f\left(x^{k+1}\right)\right\|^2 \mid [1]\right]\right]\\
&= \mathbb{E}\left[\frac{c\delta_{\max}}{G_{\widehat{C}}^k}\sum_{i\in\mathcal{G}_{\widehat{C}}^k}4\mathbb{E}_k\left[\left\|\nabla f_i\left(x^{k+1}\right)-\nabla f\left(x^{k+1}\right)\right\|^2 \mid [1]\right]\right]\\
&= \frac{\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}c\delta_{\max}}{(1-\delta_{\max})\frac{\delta n}{\delta_{\max}}}\sum_{i\in\mathcal{G}}4\mathbb{E}_k\left[\left\|\nabla f_i\left(x^{k+1}\right)-\nabla f\left(x^{k+1}\right)\right\|^2 \mid [1]\right]\\
&\overset{\text{(As.9)}}{\leqslant} \frac{4\delta n\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}c\delta_{\max}}{(1-\delta_{\max})\frac{\delta n}{\delta_{\max}}}\left(B\mathbb{E}\left[\left\|\nabla f\left(x^{k+1}\right)\right\|^2\right]+\zeta^2\right)\\
&\overset{(7)}{\leqslant} 8\frac{4\delta n\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}c\delta_{\max}}{(1-\delta_{\max})\frac{\delta n}{\delta_{\max}}}B\mathbb{E}\left[\left\|\nabla f\left(x^{k}\right)\right\|^2\right]+8\frac{4\delta n\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}c\delta_{\max}}{(1-\delta_{\max})\frac{\delta n}{\delta_{\max}}}B\mathbb{E}\left[\left\|\nabla f\left(x^{k+1}\right)-\nabla f\left(x^{k}\right)\right\|^2\right]\\
&+4\frac{4\delta n\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}c\delta_{\max}}{(1-\delta_{\max})\frac{\delta n}{\delta_{\max}}}\zeta^2\\
&\leqslant \frac{8\delta n\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}c\delta_{\max}}{(1-\delta_{\max})\frac{\delta n}{\delta_{\max}}}B\mathbb{E}\left[\left\|\nabla f\left(x^{k}\right)\right\|^2\right]+\frac{8\delta n\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}c\delta_{\max}}{(1-\delta_{\max})\frac{\delta n}{\delta_{\max}}}BL^2\mathbb{E}\left[\left\|x^{k+1}-x^k\right\|^2\right]+4\frac{\delta n\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}c\delta_{\max}}{(1-\delta_{\max})\frac{\delta n}{\delta_{\max}}}\zeta^2.
\end{aligned}
$$

□

Lemma G.6. *Let Assumptions 4, 5, 6 hold and the Compression Operator satisfy Definition 2.2. Also let us introduce the notation*

$$
\texttt{ARAgg}_Q^{k+1}=\texttt{ARAgg}\left(\text{clip}_{\lambda_{k+1}}\left(\mathcal{Q}\left(\widehat{\Delta}_1(x^{k+1},x^k)\right)\right),\ldots,\text{clip}_{\lambda_{k+1}}\left(\mathcal{Q}\left(\widehat{\Delta}_C(x^{k+1},x^k)\right)\right)\right).
$$

Then for all $k\geqslant 0$ the iterates produced by Byz-VR-MARINA-PP *(Algorithm 1) satisfy*

$$
\begin{aligned}
T_2 &= \mathbb{E}\left[\mathbb{E}_k\left[\left\|\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k}\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right)-\texttt{ARAgg}_Q^{k+1}\right\|^2 \mid [3]\right]\right]\\
&\leqslant 8\mathcal{P}_{\mathcal{G}_C^k}\left(10(1+\omega)\frac{\mathcal{L}_\pm^2}{b}+(10\omega+1)L_\pm^2+10\omega L^2\right)c\delta_{\max}\mathbb{E}\left[\|x^{k+1}-x^k\|^2\right],
\end{aligned}
$$

where $\mathcal{P}_{\mathcal{G}_C^k}=\text{Prob}\left\{i\in\mathcal{G}_C^k \mid G_C^k\geqslant(1-\delta_{\max})C\right\}$.

Proof. Let us consider second term, since

$$
\begin{aligned}
T_2 &= \mathbb{E}\left[\mathbb{E}_k\left[\left\|\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k}\operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right) - \texttt{ARAgg}_Q^{k+1}\right\|^2 \mid [3]\right]\right] \\
&\leqslant \mathbb{E}\left[\frac{c\delta_{\max}}{D_2}\sum_{\substack{i,l\in\mathcal{G}_C^k\\ i\neq l}}\mathbb{E}_k\left[\left\|\operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right) - \operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_l\left(x^{k+1},x^k\right)\right)\right)\right\|^2 \mid [3]\right]\right],
\end{aligned}
$$

where $D_2 = G_C^k(G_C^k - 1)$

Let us consider pair-wise differences:

$$
\begin{aligned}
T_2'(i,l) &= \mathbb{E}_k\left[\left\|\operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right) - \operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_l\left(x^{k+1},x^k\right)\right)\right)\right\|^2 \mid [3]\right] \\
&\overset{(7)}{\leqslant} 2\mathbb{E}_k\left[\left\|\operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right) - \Delta_i\left(x^{k+1},x^k\right) + \Delta_l\left(x^{k+1},x^k\right) - \operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_l\left(x^{k+1},x^k\right)\right)\right)\right\|^2 \mid [3]\right] \\
&+ 2\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right) - \Delta_l\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right] \\
&\overset{(7)}{\leqslant} 4\mathbb{E}_k\left[\left\|\operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right) - \Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right] \\
&+ 4\mathbb{E}_k\left[\left\|\Delta_l\left(x^{k+1},x^k\right) - \operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_l\left(x^{k+1},x^k\right)\right)\right)\right\|^2 \mid [3]\right] \\
&+ 2\mathbb{E}_k\left[\left\|\Delta_l\left(x^{k+1},x^k\right) - \Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]] \\
&\overset{(7)}{\leqslant} 4\mathbb{E}_k\left[\left\|\operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right) - \Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right] \\
&+ 4\mathbb{E}_k\left[\left\|\Delta_l\left(x^{k+1},x^k\right) - \operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_l\left(x^{k+1},x^k\right)\right)\right)\right\|^2 \mid [3]\right] \\
&+ 4\mathbb{E}_k\left[\left\|\Delta_l\left(x^{k+1},x^k\right) - \Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right] \\
&+ 4\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right) - \Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right].
\end{aligned}
$$

Now we can combine all parts together:

$$
\begin{aligned}
\widehat{T}_2 &= \mathbb{E}\left[\frac{1}{G_C^k(G_C^k-1)}\sum_{\substack{i,l\in\mathcal{G}_C^k\\ i\neq l}} T_2'(i,l)\right]\\
&\leqslant \mathbb{E}\left[\frac{1}{D_2}\sum_{\substack{i,l\in\mathcal{G}_C^k\\ i\neq l}} 4\mathbb{E}_k\left[\left\|\operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right]\\
&+\mathbb{E}\left[\frac{1}{D_2}\sum_{\substack{i,l\in\mathcal{G}_C^k\\ i\neq l}} 4\mathbb{E}_k\left[\left\|\Delta_l\left(x^{k+1},x^k\right)-\operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_l\left(x^{k+1},x^k\right)\right)\right)\right\|^2 \mid [3]\right]\right]\\
&+\mathbb{E}\left[\frac{1}{D_2}\sum_{\substack{i,l\in\mathcal{G}_C^k\\ i\neq l}} 4\mathbb{E}_k\left[\left\|\Delta_l\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right]\\
&+\mathbb{E}\left[\frac{1}{D_2}\sum_{\substack{i,l\in\mathcal{G}_C^k\\ i\neq l}} 4\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right].
\end{aligned}
$$

Combining terms together we have

$$
\begin{aligned}
\widehat{T}_2 &\leqslant \mathbb{E}\left[\frac{1}{D_2}\sum_{\substack{i,l\in\mathcal{G}_C^k\\ i\neq l}} 8\mathbb{E}_k\left[\left\|\operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right]\\
&+\mathbb{E}\left[\frac{1}{D_2}\sum_{\substack{i,l\in\mathcal{G}_C^k\\ i\neq l}} 8\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right].
\end{aligned}
$$

It leads to

$$
\begin{aligned}
\widehat{T}_2 &\leqslant \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 8\mathbb{E}_k\left[\left\|\operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right]\\
&+\mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 8\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right]\\
&\leqslant \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 80\mathbb{E}_k\left[\left\|\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right]\\
&+\mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 8\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right].
\end{aligned}
$$

Using variance decomposition we get

$$\begin{aligned}
\widehat{T}_2 \leqslant\ & \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 80\mathbb{E}_k\left[\left\|\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right\|^2 \mid [3]\right]\right] \\
& - \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 80\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
& + \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 8\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right].
\end{aligned}$$

Using properties of unbiased compressors (Definition 2.2) we have

$$\begin{aligned}
\widehat{T}_2 \leqslant\ & \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 80(1+\omega)\mathbb{E}_k\left[\left\|\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
& - \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 80\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
& + \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 8\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right].
\end{aligned}$$

Also we have

$$\begin{aligned}
\widehat{T}_2 \leqslant\ & \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 80(1+\omega)\mathbb{E}_k\left[\left\|\widehat{\Delta}_i\left(x^{k+1},x^k\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
& + \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 80(1+\omega)\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
& - \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 80\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
& + \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 8\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right].
\end{aligned}$$

Let us simplify the inequality:

$$\begin{aligned}
\widehat{T}_2 \leqslant\ & \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 80(1+\omega)\mathbb{E}_k\left[\left\|\widehat{\Delta}_i\left(x^{k+1},x^k\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
& + \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 80\omega\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
& + \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 8\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right].
\end{aligned}$$

Using decomposition we have

$$\begin{aligned}\widehat{T}_2 \leqslant\ & \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 80(1+\omega)\mathbb{E}_k\left[\left\|\widehat{\Delta}_i\left(x^{k+1},x^k\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right]\\ &+\mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 80\omega\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right]\\ &+\mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 8\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right]\\ &+\mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 80\omega\mathbb{E}_k\left[\left\|\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right].\end{aligned}$$

Using similar argument in previous lemma we obtain

$$\begin{aligned}\widehat{T}_2 \leqslant\ & \mathbb{E}\left[\frac{\mathcal{P}_{\mathcal{G}_C^k}}{G}\sum_{i\in\mathcal{G}} 80(1+\omega)\mathbb{E}_k\left[\left\|\widehat{\Delta}_i\left(x^{k+1},x^k\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right]\\ &+\mathbb{E}\left[\frac{\mathcal{P}_{\mathcal{G}_C^k}}{G}\sum_{i\in\mathcal{G}} 80\omega\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right]\\ &+\mathbb{E}\left[\frac{\mathcal{P}_{\mathcal{G}_C^k}}{G}\sum_{i\in\mathcal{G}} 8\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right]\\ &+\mathbb{E}\left[\frac{\mathcal{P}_{\mathcal{G}_C^k}}{G}\sum_{i\in\mathcal{G}} 80\omega\mathbb{E}_k\left[\left\|\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right].\end{aligned}$$

Using Assumptions 4, 5, 6:

$$\begin{aligned}\widehat{T}_2 \leqslant\ & \mathbb{E}\left[80(1+\omega)\mathcal{P}_{\mathcal{G}_C^k}\frac{\mathcal{L}_\pm^2}{b}\|x^{k+1}-x^k\|^2\right]\\ &+\mathbb{E}\left[8(10\omega+1)\mathcal{P}_{\mathcal{G}_C^k}L_\pm^2\|x^{k+1}-x^k\|^2\right]\\ &+\mathbb{E}\left[80\mathcal{P}_{\mathcal{G}_C^k}\omega L^2\|x^{k+1}-x^k\|^2\right].\end{aligned}$$

Finally, we obtain

$$\begin{aligned}T_2 &= \mathbb{E}\left[\mathbb{E}_k\left[\left\|\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k}\operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right)-\texttt{ARAgg}_Q^{k+1}\right\|^2 \mid [3]\right]\right]\\ &\leqslant 8\mathcal{P}_{\mathcal{G}_C^k}\left(10(1+\omega)\frac{\mathcal{L}_\pm^2}{b}+(10\omega+1)L_\pm^2+10\omega L^2\right)c\delta_{\max}\mathbb{E}\left[\|x^{k+1}-x^k\|^2\right].\end{aligned}$$

□

Lemma G.7. *Let Assumptions 1 and 4 hold. Also let us introduce the notation*

$$\texttt{ARAgg}_Q^{k+1} = \texttt{ARAgg}\left(\operatorname{clip}_{\lambda_{k+1}}\left(\mathcal{Q}\left(\widehat{\Delta}_1(x^{k+1},x^k)\right)\right),\ldots,\operatorname{clip}_{\lambda_{k+1}}\left(\mathcal{Q}\left(\widehat{\Delta}_C(x^{k+1},x^k)\right)\right)\right).$$

Assume that $\lambda_{k+1}=\alpha_{\lambda_{k+1}}\|x^{k+1}-x^k\|$. Then for all $k\geqslant 0$ the iterates produced by Byz-VR-MARINA-PP *(Algorithm 1) satisfy*

$$\begin{aligned}T_3 &= \mathbb{E}\left[\mathbb{E}_k\left[\left\|\nabla f(x^{k+1})-\nabla f(x^k)-\texttt{ARAgg}_Q^{k+1}\right\|^2 \mid [2]\right]\right]\\ &\leqslant 2(L^2+F_{\mathcal{A}}^2\alpha_{\lambda_{k+1}}^2)\mathbb{E}\left[\left\|x^{k+1}-x^k\right\|^2\right]\end{aligned}$$

Proof.

$$
\begin{aligned}
T_3 &= \mathbb{E}\left[\mathbb{E}_k\left[\left\|\nabla f(x^{k+1}) - \nabla f(x^k) - \texttt{ARAgg}_Q^{k+1}\right\|^2 \mid [2]\right]\right] \\
&\overset{(7)}{\leqslant} \mathbb{E}\left[\mathbb{E}_k\left[2\left\|\nabla f(x^{k+1}) - \nabla f(x^k)\right\|^2 + 2\left\|\texttt{ARAgg}_Q^{k+1}\right\|^2 \mid [2]\right]\right]
\end{aligned}
$$

Using L-smoothness and Assumption 1 we have

$$
\begin{aligned}
T_3 &\overset{(7)}{\leqslant} \mathbb{E}\left[\mathbb{E}_k\left[2L^2\left\|x^{k+1} - x^k\right\|^2 + 2F_{\mathcal{A}}^2\lambda_{k+1}^2 \mid [2]\right]\right] \\
&\leqslant \mathbb{E}\left[\mathbb{E}_k\left[2L^2\left\|x^{k+1} - x^k\right\|^2 + 2F_{\mathcal{A}}^2\alpha_{\lambda_{k+1}}^2\|x^{k+1} - x^k\|^2 \mid [2]\right]\right] \\
&\leqslant 2(L^2 + F_{\mathcal{A}}^2\alpha_{\lambda_{k+1}}^2)\mathbb{E}\left[\left\|x^{k+1} - x^k\right\|^2\right].
\end{aligned}
$$

□

Lemma G.8. *Let Assumptions 1, 4, 5, 6, 9 hold and Compression Operator satisfy Definition 2.2. Also let us introduce the notation*

$$
\texttt{ARAgg}_Q^{k+1} = \texttt{ARAgg}\left(\operatorname{clip}_{\lambda_{k+1}}\left(\mathcal{Q}\left(\widehat{\Delta}_1(x^{k+1}, x^k)\right)\right), \ldots, \operatorname{clip}_{\lambda_{k+1}}\left(\mathcal{Q}\left(\widehat{\Delta}_C(x^{k+1}, x^k)\right)\right)\right).
$$

Then for all $k \geqslant 0$ the iterates produced by Byz-VR-MARINA-PP *(Algorithm 1) satisfy*

$$
\begin{aligned}
\mathbb{E}\left[\left\|g^{k+1} - \nabla f\left(x^{k+1}\right)\right\|^2\right] \leqslant \left(1 - \frac{p}{4}\right)\mathbb{E}\left[\left\|g^k - \nabla f\left(x^k\right)\right\|^2\right] \\
+ 24c\delta B\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] + 12c\delta\zeta^2 + \frac{pA}{4}\|x^{k+1} - x^k\|^2,
\end{aligned}
$$

where

$$
\begin{aligned}
A = \frac{4}{p}&\left(\frac{80}{p}\frac{p_G\mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}\omega + 24c\delta B + \frac{4}{p}(1-p_G) + \frac{160}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}c\delta_{\max}\omega\right)L^2 \\
+ \frac{4}{p}&\left(\frac{8}{p}\frac{p_G\mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}(10\omega+1) + \frac{16}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}c\delta_{\max}(10\omega+1)\right)L_{\pm}^2 \\
+ \frac{4}{p}&\left(\frac{160}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}(1+\omega)c\delta_{\max} + \frac{80}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}(1+\omega)\frac{(1-\delta)n}{C(1-\delta_{\max})}\right)\frac{\mathcal{L}_{\pm}^2}{b} \\
+ \frac{4}{p}&\left(\frac{4}{p}(1-p_G)F_{\mathcal{A}}^2\alpha_{\lambda_{k+1}}^2\right),
\end{aligned}
$$

and where $p_G = \operatorname{Prob}\left\{G_C^k \geqslant (1-\delta_{\max})C\right\}$ and $\mathcal{P}_{\mathcal{G}_C^k} = \operatorname{Prob}\left\{i \in \mathcal{G}_C^k \mid G_C^k \geqslant (1-\delta_{\max})\,C\right\}$.

Proof. Let us combine bounds for A_1 and A_2 together:

$$
\begin{aligned}
A_0 &= \mathbb{E}\left[\left\|g^{k+1}-\nabla f\left(x^{k+1}\right)\right\|^2\right] \\
&\leqslant \left(1+\frac{p}{2}\right)\mathbb{E}\left[\left\|\overline{g}^{k+1}-\nabla f\left(x^{k+1}\right)\right\|^2\right]+\left(1+\frac{2}{p}\right)\mathbb{E}\left[\left\|g^{k+1}-\overline{g}^{k+1}\right\|^2\right] \\
&\leqslant \left(1+\frac{p}{2}\right)A_1+\left(1+\frac{2}{p}\right)A_2 \\
&\leqslant \left(1+\frac{p}{2}\right)(1-p)\left(1+\frac{p}{4}\right)\mathbb{E}\left[\left\|g^k-\nabla f\left(x^k\right)\right\|^2\right] \\
&+ \left(1+\frac{p}{2}\right)(1-p)p_G\left(1+\frac{4}{p}\right)\frac{2\cdot\mathcal{P}_{\mathcal{G}^k_C}(1-\delta)n}{C(1-\delta_{\max})}\left(10\omega L^2+(10\omega+1)L_\pm^2+\frac{10(\omega+1)\mathcal{L}_\pm^2}{b}\right)\mathbb{E}\left[]\|x^{k+1}-x^k\|^2\right] \\
&+ \left(1+\frac{p}{2}\right)p\left(\frac{\delta_{\max}\cdot\mathcal{P}_{\mathcal{G}^k_{\widehat{C}}}}{(1-\delta_{\max})}\mathbb{E}\left[B\|\nabla f(x)\|^2+\zeta^2\right]\right) \\
&+ \left(1+\frac{2}{p}\right)p\mathbb{E}\left[\mathbb{E}_k\left[\left\|\texttt{ARAgg}\left(\nabla f_1(x^{k+1}),\ldots,\nabla f_{\widehat{C}}(x^{k+1})\right)-\nabla f(x^{k+1})\right\|^2\right]\mid[1]\right] \\
&+ \left(1+\frac{2}{p}\right)(1-p)p_G\mathbb{E}\left[\mathbb{E}_k\left[\left\|\frac{1}{G^k_c}\sum_{i\in\mathcal{G}^k_C}\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right)-\texttt{ARAgg}^{k+1}_Q\right\|^2\mid[3]\right]\right] \\
&+ \left(1+\frac{2}{p}\right)(1-p)(1-p_G)\mathbb{E}\left[\mathbb{E}_k\left[\left\|\nabla f(x^{k+1})-\nabla f(x^k)-\texttt{ARAgg}^{k+1}_Q\right\|^2\mid[2]\right]\right].
\end{aligned}
$$

Finally, we obtain the following bound:

$$
\begin{aligned}
A_0 &\overset{(7)}{\leqslant} \left(1-\frac{p}{4}\right)\mathbb{E}\left[\left\|g^k-\nabla f\left(x^k\right)\right\|^2\right] \\
&+ \frac{8}{p}\frac{\mathcal{P}_{\mathcal{G}^k_C}(1-\delta)n}{C(1-\delta_{\max})}p_G\left(10\omega L^2+(10\omega+1)L_\pm^2+\frac{10(\omega+1)\mathcal{L}_\pm^2}{b}\right)\mathbb{E}\left[\|x^{k+1}-x^k\|^2\right] \\
&+ 2p\left(\frac{\delta_{\max}\cdot\mathcal{P}_{\mathcal{G}^k_{\widehat{C}}}}{(1-\delta_{\max})}\mathbb{E}\left[B\|\nabla f(x)\|^2+\zeta^2\right]\right) \\
&+ (p+2)\,\mathbb{E}\left[\mathbb{E}_k\left[\left\|\texttt{ARAgg}\left(\nabla f_1(x^{k+1}),\ldots,\nabla f_n(x^{k+1})\right)-\nabla f(x^{k+1})\right\|^2\right]\mid[1]\right] \\
&+ \frac{2}{p}p_G\mathbb{E}\left[\mathbb{E}_k\left[\left\|\frac{1}{G^k_c}\sum_{i\in\mathcal{G}^k_C}\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right)-\texttt{ARAgg}^{k+1}_Q\right\|^2\mid[3]\right]\right] \\
&+ \frac{2}{p}(1-p_G)\mathbb{E}\left[\mathbb{E}_k\left[\left\|\nabla f(x^{k+1})-\nabla f(x^k)-\texttt{ARAgg}^{k+1}_Q\right\|^2\mid[2]\right]\right]
\end{aligned}
$$

Now we can apply Lemmas G.5, G.6, G.7 we have

$$
\begin{aligned}
A_0 &= \mathbb{E}\left[\left\|g^{k+1} - \nabla f\left(x^{k+1}\right)\right\|^2\right] \\
&\leqslant \left(1 - \frac{p}{4}\right)\mathbb{E}\left[\left\|g^k - \nabla f\left(x^k\right)\right\|^2\right] \\
&+ \frac{8}{p}\frac{\mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}p_G\left(10\omega L^2 + (10\omega+1)L_\pm^2 + \frac{10(\omega+1)\mathcal{L}_\pm^2}{b}\right)\mathbb{E}\left[\|x^{k+1} - x^k\|^2\right] \\
&+ (p+2)\left(\frac{8\delta n\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}c\delta_{\max}}{(1-\delta_{\max})\frac{\delta n}{\delta_{\max}}}B\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] + \frac{8\delta n\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}c\delta_{\max}}{(1-\delta_{\max})\frac{\delta n}{\delta_{\max}}}BL^2\mathbb{E}\left[\left\|x^{k+1} - x^k\right\|^2\right]\right) \\
&+ 4\left(p+2\right)\frac{\delta n\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}c\delta_{\max}}{(1-\delta_{\max})\frac{\delta n}{\delta_{\max}}}\zeta^2 \\
&+ \frac{2}{p}p_G\mathbb{E}\left[80(1+\omega)\mathcal{P}_{\mathcal{G}_C^k}\frac{\mathcal{L}_\pm^2}{b}c\delta_{\max}\|x^{k+1} - x^k\|^2\right] \\
&+ \frac{2}{p}p_G\mathbb{E}\left[8(10\omega+1)\mathcal{P}_{\mathcal{G}_C^k}L_\pm^2c\delta_{\max}\|x^{k+1} - x^k\|^2\right] \\
&+ \frac{2}{p}p_G\mathbb{E}\left[80\mathcal{P}_{\mathcal{G}_C^k}\omega L^2c\delta_{\max}\|x^{k+1} - x^k\|^2\right] \\
&+ \frac{2}{p}(1-p_G)2(L^2 + F_{\mathcal{A}}^2\alpha_{\lambda_{k+1}}^2)\mathbb{E}\left[\left\|x^{k+1} - x^k\right\|^2\right] \\
&+ 2p\frac{\delta_{\max}\cdot\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}}{(1-\delta_{\max})}\mathbb{E}\left[B\|\nabla f(x)\|^2 + \zeta^2\right].
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
\mathbb{E}\left[\left\|g^{k+1} - \nabla f\left(x^{k+1}\right)\right\|^2\right] \leqslant{}& \left(1 - \frac{p}{4}\right)\mathbb{E}\left[\left\|g^k - \nabla f\left(x^k\right)\right\|^2\right] \\
&+ \widehat{B}\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] + \widehat{C}\zeta^2 + \frac{pA}{4}\|x^{k+1} - x^k\|^2,
\end{aligned}
$$

where

$$
\begin{aligned}
A = {}& \frac{32p_G}{p^2}\frac{\mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}\left(10\omega L^2 + (10\omega+1)L_\pm^2 + \frac{10(\omega+1)\mathcal{L}_\pm^2}{b}\right) \\
&+ \frac{4}{p}\left(\frac{2}{p}p_G 80(1+\omega)\mathcal{P}_{\mathcal{G}_C^k}\frac{\mathcal{L}_\pm^2}{b}c\delta_{\max} + \frac{2}{p}p_G 8(10\omega+1)\mathcal{P}_{\mathcal{G}_C^k}L_\pm^2c\delta_{\max} + \frac{2}{p}p_G 80\mathcal{P}_{\mathcal{G}_C^k}\omega L^2c\delta_{\max}\right) \\
&+ \frac{4}{p}\cdot 8(p+2)\frac{\delta n\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}c\delta_{\max}}{(1-\delta_{\max})\frac{\delta n}{\delta_{\max}}}BL^2 + \frac{16(1-p_G)}{p^2}(L^2 + F_{\mathcal{A}}^2\alpha_{\lambda_{k+1}}^2),
\end{aligned}
$$

and

$$
\widehat{B} = 2\frac{\delta_{\max}\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}}{(1-\delta_{\max})}B(12c\delta_{\max}+1) \quad \widehat{C} = 2\frac{\delta_{\max}\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}}{(1-\delta_{\max})}B(6c\delta_{\max}+1)
$$

Once we simplify the equation, we obtain

$$
\begin{aligned}
A = {}& \frac{4}{p}\left(\frac{80}{p}\frac{p_G\mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}\omega + 24\frac{\delta_{\max}\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}c\delta_{\max}}{(1-\delta_{\max})}B + \frac{4}{p}(1-p_G) + \frac{160}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}c\delta_{\max}\omega\right)L^2 \\
&+ \frac{4}{p}\left(\frac{8}{p}\frac{p_G\mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}(10\omega+1) + \frac{16}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}c\delta_{\max}(10\omega+1)\right)L_\pm^2 \\
&+ \frac{4}{p}\left(\frac{160}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}(1+\omega)c\delta_{\max} + \frac{80}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}(1+\omega)\frac{(1-\delta)n}{C(1-\delta_{\max})}\right)\frac{\mathcal{L}_\pm^2}{b} \\
&+ \frac{4}{p}\left(\frac{4}{p}(1-p_G)F_{\mathcal{A}}^2\alpha_{\lambda_{k+1}}^2\right).
\end{aligned}
$$

□

Theorem G.1. *Let Assumptions 1, 4, 5, 6, 9 hold. Setting* $\lambda_{k+1} = 2\max_{i\in\mathcal{G}} L_i \left\|x^{k+1} - x^k\right\|$. *Assume that*

$$0 < \gamma \leqslant \frac{1}{L + \sqrt{A}}, \quad 4\widehat{B} < p,$$

where

$$\begin{aligned}
A = &\frac{4}{p}\left(\frac{80}{p}\frac{p_G\mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}\omega + 24\frac{\delta_{\max}\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}c\delta_{\max}}{(1-\delta_{\max})}B + \frac{4}{p}(1-p_G) + \frac{160}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}c\delta_{\max}\omega\right)L^2 \\
&+ \frac{4}{p}\left(\frac{8}{p}\frac{p_G\mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}(10\omega+1) + \frac{16}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}c\delta_{\max}(10\omega+1)\right)L_{\pm}^2 \\
&+ \frac{4}{p}\left(\frac{160}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}(1+\omega)c\delta_{\max} + \frac{80}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}(1+\omega)\frac{(1-\delta)n}{C(1-\delta_{\max})}\right)\frac{\mathcal{L}_{\pm}^2}{b} \\
&+ \frac{4}{p}\left(\frac{4}{p}(1-p_G)F_{\mathcal{A}}^2\left(\max_{i\in\mathcal{G}} L_i\right)^2\right),
\end{aligned}$$

$$\widehat{B} = 2\frac{\delta_{\max}\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}}{(1-\delta_{\max})}B(12c\delta_{\max}+1) \quad \widehat{C} = 2\frac{\delta_{\max}\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}}{(1-\delta_{\max})}B(6c\delta_{\max}+1)$$

and

$$\begin{aligned}
\mathcal{P}_{\mathcal{G}_C^k} &= \frac{C}{np_G}\cdot\sum_{(1-\delta_{\max})C\leqslant t\leqslant C}\left(\binom{G-1}{t-1}\binom{n-G}{C-t}\left(\binom{n}{C}\right)^{-1}\right), \\
p_G &= \mathbb{P}\left\{G_C^k \geqslant (1-\delta_{\max})C\right\} \\
&= \sum_{\lceil(1-\delta_{\max})C\rceil\leqslant t\leqslant C}\left(\binom{G}{t}\binom{n-G}{C-t}\binom{n-1}{C-1}^{-1}\right).
\end{aligned}$$

Then for all $K \geqslant 0$ *the iterates produced by Byz-VR-MARINA (Algorithm 1) satisfy*

$$\mathbb{E}\left[\left\|\nabla f\left(\widehat{x}^K\right)\right\|^2\right] \leqslant \frac{2\Phi^0}{\gamma\left(1-\frac{4\widehat{B}}{p}\right)(K+1)} + \frac{4\widehat{C}\zeta^2}{p-4\widehat{B}},$$

where $\widehat{x}^K$ *is choosen uniformly at random from* $x^0, x^1, \ldots, x^K$, *and* $\Phi^0 = f\left(x^0\right) - f^* + \frac{2\gamma}{p}\left\|g^0 - \nabla f\left(x^0\right)\right\|^2$.

Proof of Theorem G.1. For all $k \geqslant 0$ we introduce $\Phi^k = f\left(x^k\right) - f^* + \frac{2\gamma}{p}\left\|g^k - \nabla f\left(x^k\right)\right\|^2$. Using the results of Lemmas G.8 and G.2, we derive

$$\begin{aligned}
\mathbb{E}\left[\Phi^{k+1}\right] &\overset{(G.2)}{\leqslant} \mathbb{E}\left[f\left(x^k\right) - f^* - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\left\|x^{k+1} - x^k\right\|^2 + \frac{\gamma}{2}\left\|g^k - \nabla f\left(x^k\right)\right\|^2\right] \\
&- \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] + \frac{2\gamma}{p}\mathbb{E}\left[\left\|g^{k+1} - \nabla f\left(x^{k+1}\right)\right\|^2\right] \\
&\overset{(G.8)}{\leqslant} \mathbb{E}\left[f\left(x^k\right) - f^* - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\left\|x^{k+1} - x^k\right\|^2 + \frac{\gamma}{2}\left\|g^k - \nabla f\left(x^k\right)\right\|^2\right] \\
&- \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] + \frac{2\gamma}{p}\left(1 - \frac{p}{4}\right)\mathbb{E}\left[\left\|g^k - \nabla f\left(x^k\right)\right\|^2\right] \\
&+ \frac{2\gamma}{p}\left(\widehat{B}\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] + \widehat{C}\zeta^2 + \frac{pA}{4}\|x^{k+1} - x^k\|^2\right) \\
&= \mathbb{E}\left[f\left(x^k\right) - f^*\right] + \frac{2\gamma}{p}\left(\left(1 - \frac{p}{4}\right) + \frac{p}{4}\right)\mathbb{E}\left[\left\|g^k - \nabla f\left(x^k\right)\right\|^2\right] + \frac{2\widehat{C}\zeta^2\gamma}{p} \\
&+ \frac{1}{2\gamma}\left(1 - L\gamma - A\gamma^2\right)\mathbb{E}\left[\|x^{k+1} - x^k\|^2\right] - \frac{\gamma}{2}\left(1 - \frac{4\widehat{B}}{p}\right)\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] \\
&= \mathbb{E}\left[\Phi^k\right] + \frac{2\widehat{C}\zeta^2\gamma}{p} + \frac{1}{2\gamma}\left(1 - L\gamma - A\gamma^2\right)\mathbb{E}\left[\|x^{k+1} - x^k\|^2\right] \\
&- \frac{\gamma}{2}\left(1 - \frac{4\widehat{B}}{p}\right)\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right].
\end{aligned}$$

Using choice of stepsize and second condition: $0 < \gamma \leqslant \frac{1}{L+\sqrt{A}}, 4\widehat{B} < p$ and lemma C.1 we have

$$\mathbb{E}\left[\Phi^{k+1}\right] \leqslant \mathbb{E}\left[\Phi^k\right] + \frac{2\widehat{C}\zeta^2\gamma}{p} - \frac{\gamma}{2}\left(1 - \frac{4\widehat{B}}{p}\right)\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right]$$

Next, we have $\frac{\gamma}{2}\left(1 - \frac{4\widehat{B}}{p}\right) > 0$ and $\Phi^{k+1} \geqslant 0$. Therefore, summing up the above inequality for $k = 0, 1, \ldots, K$ and rearranging the terms, we get

$$\begin{aligned}
\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] &\leqslant \frac{2}{\gamma\left(1 - \frac{4\widehat{B}}{p}\right)(K+1)}\sum_{k=0}^{K}\left(\mathbb{E}\left[\Phi^k\right] - \mathbb{E}\left[\Phi^{k+1}\right]\right) \\
&+ \frac{4\widehat{C}\zeta^2}{p - 4\widehat{B}} \\
&= \frac{2\left(\mathbb{E}\left[\Phi^0\right] - \mathbb{E}\left[\Phi^{k+1}\right]\right)}{\gamma\left(1 - \frac{4\widehat{B}}{p}\right)(K+1)} + \frac{4\widehat{C}\zeta^2}{p - 4\widehat{B}} \\
&\leqslant \frac{2\mathbb{E}\left[\Phi^0\right]}{\gamma\left(1 - \frac{4\widehat{B}}{p}\right)(K+1)} + \frac{4\widehat{C}\zeta^2}{p - 4\widehat{B}}.
\end{aligned}$$

□

Theorem G.2. *Let Assumptions 1, 4, 5, 6, 9, 8 hold. Setting* $\lambda_{k+1} = \max_{i\in\mathcal{G}} L_i \left\|x^{k+1} - x^k\right\|$. *Assume that*

$$0 < \gamma \leqslant \min\left\{\frac{1}{L + \sqrt{2A}}\right\}, \quad 8\widehat{B} < p$$

where

$$A = \frac{4}{p}\left(\frac{80}{p}\frac{p_G \mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}\omega + 24\frac{\delta_{\max}\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}c\delta_{\max}}{(1-\delta_{\max})}B + \frac{4}{p}(1-p_G) + \frac{160}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}c\delta_{\max}\omega\right)L^2$$
$$+ \frac{4}{p}\left(\frac{8}{p}\frac{p_G\mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}(10\omega+1) + \frac{16}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}c\delta_{\max}(10\omega+1)\right)L_{\pm}^2$$
$$+ \frac{4}{p}\left(\frac{160}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}(1+\omega)c\delta_{\max} + \frac{80}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}(1+\omega)\frac{(1-\delta)n}{C(1-\delta_{\max})}\right)\frac{\mathcal{L}_{\pm}^2}{b}$$
$$+ \frac{4}{p}\left(\frac{4}{p}(1-p_G)F_{\mathcal{A}}^2\left(\max_{i\in\mathcal{G}}L_i\right)^2\right),$$

$$\widehat{B} = 2\frac{\delta_{\max}\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}}{(1-\delta_{\max})}B(12c\delta_{\max}+1) \quad \widehat{C} = 2\frac{\delta_{\max}\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}}{(1-\delta_{\max})}B(6c\delta_{\max}+1),$$

and where $p_G = \text{Prob}\left\{G_C^k \geqslant (1-\delta_{\max})C\right\}$ *and* $\mathcal{P}_{\mathcal{G}_C^k} = \text{Prob}\left\{i \in \mathcal{G}_C^k \mid G_C^k \geqslant (1-\delta_{\max})\,C\right\}$. *Then for all* $K \geqslant 0$ *the iterates produced by Byz-VR-MARINA (Algorithm 1) satisfy*

$$\mathbb{E}\left[f\left(x^K\right) - f\left(x^*\right)\right] \leqslant (1-\rho)^K\,\Phi^0 + \frac{4\widehat{C}\gamma\zeta^2}{p\rho},$$

where $\rho = \min\left[\gamma\mu\left(1-\frac{8\widehat{B}}{p}\right), \frac{p}{8}\right]$ *and* $\Phi^0 = f\left(x^0\right) - f^* + \frac{4\gamma}{p}\left\|g^0 - \nabla f\left(x^0\right)\right\|^2$.

Proof. For all $k \geqslant 0$ we introduce $\Phi^k = f\left(x^k\right) - f^* + \frac{4\gamma}{p}\left\|g^k - \nabla f\left(x^k\right)\right\|^2$. Using the results of Lemmas G.8 and G.2, we derive

$$\begin{aligned}
\mathbb{E}\left[\Phi^{k+1}\right] &\overset{(G.2)}{\leqslant} \mathbb{E}\left[f\left(x^k\right) - f^* - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\left\|x^{k+1}-x^k\right\|^2 + \frac{\gamma}{2}\left\|g^k - \nabla f\left(x^k\right)\right\|^2\right] \\
&- \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] + \frac{4\gamma}{p}\mathbb{E}\left[\left\|g^{k+1} - \nabla f\left(x^{k+1}\right)\right\|^2\right] \\
&\overset{(G.8)}{\leqslant} \mathbb{E}\left[f\left(x^k\right) - f^* - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\left\|x^{k+1}-x^k\right\|^2 + \frac{\gamma}{2}\left\|g^k - \nabla f\left(x^k\right)\right\|^2\right] \\
&- \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] + \frac{4\gamma}{p}\left(1-\frac{p}{4}\right)\mathbb{E}\left[\left\|g^k - \nabla f\left(x^k\right)\right\|^2\right] \\
&+ \frac{4\gamma}{p}\left(\widehat{B}\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] + \widehat{C}\zeta^2 + \frac{pA}{4}\|x^{k+1}-x^k\|^2\right) \\
&= \mathbb{E}\left[f\left(x^k\right) - f^*\right] + \frac{4\gamma}{p}\left(\left(1-\frac{p}{4}\right) + \frac{p}{8}\right)\mathbb{E}\left[\left\|g^k - \nabla f\left(x^k\right)\right\|^2\right] + \frac{4\widehat{C}\zeta^2\gamma}{p} \\
&+ \frac{1}{2\gamma}\left(1 - L\gamma - 2A\gamma^2\right)\mathbb{E}\left[\|x^{k+1}-x^k\|^2\right] - \frac{\gamma}{2}\left(1-\frac{8\widehat{B}}{p}\right)\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right].
\end{aligned}$$

Using Assumption 8 we obtain

$$\begin{aligned}
\mathbb{E}\left[\Phi^{k+1}\right] &\leqslant \mathbb{E}\left[f\left(x^k\right) - f^*\right] + \left(1-\frac{p}{8}\right)\frac{4\gamma}{p}\mathbb{E}\left[\left\|g^k - \nabla f\left(x^k\right)\right\|^2\right] + \frac{4c\delta\zeta^2\gamma}{p} \\
&+ \frac{1}{2\gamma}\left(1 - L\gamma - 2A\gamma^2\right)\mathbb{E}\left[\|x^{k+1}-x^k\|^2\right] \\
&- \gamma\mu\left(1-\frac{8\widehat{B}}{p}\right)\mathbb{E}\left[f\left(x^k\right) - f^*\right].
\end{aligned}$$

Finally, we have

$$\mathbb{E}\left[\Phi^{k+1}\right] \leqslant \left(1 - \min\left[\gamma\mu\left(1 - \frac{\widehat{B}}{p}\right), \frac{p}{8}\right]\right) \mathbb{E}\left[\Phi^{k}\right] + \frac{4\widehat{C}\zeta^2\gamma}{p}.$$

Unrolling the recurrence with $\rho = \min\left[\gamma\mu\left(1 - \frac{8\widehat{B}}{p}\right), \frac{p}{8}\right]$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\Phi^{k}\right] &\leqslant (1-\rho)^K \mathbb{E}\left[\Phi^{0}\right] + \frac{4\widehat{C}\zeta^2\gamma}{p} \sum_{k=0}^{K-1} (1-\rho)^k \\
&\leqslant (1-\rho)^K \mathbb{E}\left[\Phi^{0}\right] + \frac{4\widehat{C}\zeta^2\gamma}{p} \sum_{k=0}^{\infty} (1-\rho)^k \\
&= (1-\rho)^K \mathbb{E}\left[\Phi^{0}\right] + \frac{4\widehat{C}\gamma\zeta^2}{p\rho}
\end{aligned}$$

Taking into account $\Phi^k \geqslant f\left(x^k\right) - f\left(x^*\right)$, we get the result.

□

H Analysis for Bounded Compressors

Lemma H.1. *Let Assumptions 4, 5, 6 and 2 hold and the Compression Operator satisfy Definition 2.2. We set $\lambda_{k+1} = D_Q \max_{i,j} L_{i,j}$. Let us define "ideal" estimator:*

$$\overline{g}^{k+1} = \begin{cases} \frac{1}{G_c^k} \sum\limits_{i \in \mathcal{G}_C^k} f_i\left(x^{k+1}\right), & c_n = 1, & [1] \\ g^k + \nabla f\left(x^{k+1}\right) - \nabla f\left(x^k\right), & c_n = 0 \text{ and } G_c^k < (1-\delta_{\max})C, & [2] \\ g^k + \frac{1}{G_c^k} \sum\limits_{i \in \mathcal{G}_C^k} \operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1}, x^k\right)\right)\right), & c_n = 0 \text{ and } G_c^k \geqslant (1-\delta_{\max})C. & [3] \end{cases}$$

Then for all $k \geqslant 0$ the iterates produced by Byz-VR-MARINA-PP *(Algorithm 1) satisfy*

$$A_1 \leqslant (1-p)\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] + (1-p)p_G \frac{\mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}\left(\omega L^2 + (\omega+1)L_\pm^2 + \frac{(\omega+1)\mathcal{L}_\pm^2}{b}\right)\mathbb{E}\left[\|x^{k+1} - x^k\|^2\right].$$

where $p_G = \operatorname{Prob}\left\{G_C^k \geqslant (1-\delta_{\max})C\right\}$ and $\mathcal{P}_{\mathcal{G}_C^k} = \operatorname{Prob}\left\{i \in \mathcal{G}_C^k \mid G_C^k \geqslant (1-\delta_{\max})\,C\right\}$.

Proof. Similarly to general analysis we start from conditional expectations:

$$\begin{aligned} A_1 &= \mathbb{E}\left[\left\|\overline{g}^{k+1} - \nabla f\left(x^{k+1}\right)\right\|^2\right] \\ &= \mathbb{E}\left[\mathbb{E}_k\left[\left\|\overline{g}^{k+1} - \nabla f\left(x^{k+1}\right)\right\|^2\right]\right] \\ &= (1-p)\,p_G\mathbb{E}\left[\mathbb{E}_k\left[\left\|g^k + \frac{1}{G_c^k}\sum_{i \in \mathcal{G}_C^k} \operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1}, x^k\right)\right)\right) - \nabla f\left(x^{k+1}\right)\right\|^2\right] \mid [3]\right] \\ &\quad + (1-p)(1-p_G)\mathbb{E}\left[\mathbb{E}_k\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \mid [2]\right] + p\left\|\frac{1}{G_c^k}\sum_{i \in \mathcal{G}_C^k} f_i\left(x^{k+1}\right) - \nabla f(x^{k+1})\right\|^2. \end{aligned}$$

Using (7) and $\nabla f\left(x^k\right) - \nabla f\left(x^k\right) = 0$ we obtain

$$\begin{aligned} B_1 &= \mathbb{E}\left[\mathbb{E}_k\left[\left\|g^k + \frac{1}{G_c^k}\sum_{i \in \mathcal{G}_C^k} \operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1}, x^k\right)\right)\right) - \nabla f\left(x^{k+1}\right)\right\|^2\right] \mid [3]\right] \\ &= \mathbb{E}\left[\mathbb{E}_k\left[\left\|g^k + \frac{1}{G_c^k}\sum_{i \in \mathcal{G}_C^k} \operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1}, x^k\right)\right)\right) - \nabla f\left(x^{k+1}\right) + \nabla f\left(x^k\right) - \nabla f\left(x^k\right)\right\|^2\right] \mid [3]\right] \end{aligned}$$

Using $\lambda_{k+1} = D_Q \max_{i,j} L_{i,j}\|x^{k+1} - x^k\|$ we can guarantee that clipping operator becomes identical since we have

$$\begin{aligned} \left\|\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1}, x^k\right)\right)\right\| &\leqslant D_Q\left\|\widehat{\Delta}_i\left(x^{k+1}, x^k\right)\right\| \\ &\leqslant D_Q\left\|\frac{1}{b}\sum_{j \in m} \nabla f_{i,j}(x^{k+1}) - \nabla f_{i,j}(x^k)\right\| \\ &\leqslant D_Q \frac{1}{b}\sum_{j \in m}\left\|\nabla f_{i,j}(x^{k+1}) - \nabla f_{i,j}(x^k)\right\| \\ &\leqslant D_Q \max_j L_{i,j}\left\|x^{k+1} - x^k\right\| \\ &\leqslant D_Q \max_{i,j} L_{i,j}\left\|x^{k+1} - x^k\right\|. \end{aligned}$$

Now we have

$$B_1 = \mathbb{E}\left[\mathbb{E}_k\left[\left\|g^k + \frac{1}{G_c^k}\sum_{i\in\mathcal{G}_C^k}\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right) - \nabla f\left(x^{k+1}\right)\right\|^2\right] \mid [3]\right]$$
$$= \mathbb{E}\left[\mathbb{E}_k\left[\left\|g^k + \frac{1}{G_c^k}\sum_{i\in\mathcal{G}_C^k}\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right) - \nabla f\left(x^{k+1}\right) + \nabla f\left(x^k\right) - \nabla f\left(x^k\right)\right\|^2\right] \mid [3]\right].$$

In case without clipping we can avoid Young's inequality and obtain

$$\begin{aligned}
B_1 &\leqslant \mathbb{E}\left[\left\|g^k - \nabla f\left(x^k\right)\right\|^2\right] \\
&+ \mathbb{E}\left[\mathbb{E}_k\left[\left\|\frac{1}{G_c^k}\sum_{i\in\mathcal{G}_C^k}\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right) - \left(\nabla f(x^{k+1}) - \nabla f(x^k)\right)\right\|^2\right] \mid [3]\right] \\
&\leqslant \mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \\
&+ \mathbb{E}\left[\mathbb{E}_k\left[\left\|\frac{1}{G_c^k}\sum_{i\in\mathcal{G}_C^k}\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right) - \Delta\left(x^{k+1},x^k\right)\right\|^2\right] \mid [3]\right].
\end{aligned}$$

Let us consider last part of the inequality. Note that $G_c^k \geqslant (1-\delta_{\max})C$ in this case

$$\begin{aligned}
B_1' &= \mathbb{E}\left[\mathbb{E}_k\left[\left\|\frac{1}{G_c^k}\sum_{i\in\mathcal{G}_C^k}\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right) - \Delta\left(x^{k+1},x^k\right)\right\|^2\right] \mid [3]\right] \\
&= \mathbb{E}\left[\mathbb{E}_{S_k}\left[\mathbb{E}_k\left[\left\|\frac{1}{G_c^k}\sum_{i\in\mathcal{G}_C^k}\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right) - \Delta\left(x^{k+1},x^k\right)\right\|^2\right] \mid [3]\right]\right] \\
&\leqslant \frac{1}{C(1-\delta_{\max})}\mathbb{E}\left[\mathbb{E}_{S_k}\left[\sum_{i\in\mathcal{G}_C^k}\mathbb{E}_k\left[\left\|\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right) - \Delta\left(x^{k+1},x^k\right)\right\|^2\right] \mid [3]\right]\right] \\
&\leqslant \frac{1}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\mathbb{E}_{S_k}\left[\mathcal{I}_{\mathcal{G}_C^k}\right]\mathbb{E}_k\left[\left\|\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right) - \Delta\left(x^{k+1},x^k\right)\right\|^2\right] \mid [3]\right] \\
&= \frac{1}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\mathcal{P}_{\mathcal{G}_C^k}\cdot\mathbb{E}_k\left[\left\|\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right) - \Delta\left(x^{k+1},x^k\right)\right\|^2\right] \mid [3]\right],
\end{aligned} \tag{20}$$

where $\mathcal{I}_{\mathcal{G}_C^k}$ is an indicator function for the event $\left\{i \in \mathcal{G}_C^k \mid G_C^k \geqslant (1-\delta_{\max})\,C\right\}$ and $\mathcal{P}_{\mathcal{G}_C^k} = \text{Prob}\left\{i \in \mathcal{G}_C^k \mid G_C^k \geqslant (1-\delta_{\max})\,C\right\}$ is probability of such event. Note that $\mathbb{E}_{S_k}\left[\mathcal{I}_{\mathcal{G}_C^k}\right] = \mathcal{P}_{\mathcal{G}_C^k}$. In case of uniform sampling of clients we have

$$\begin{aligned}
\forall i \in \mathcal{G} \quad \mathcal{P}_{\mathcal{G}_C^k} &= \text{Prob}\left\{i \in \mathcal{G}_C^k \mid G_C^k \geqslant (1-\delta_{\max})\,C\right\} \\
&= \frac{C}{n}\frac{1}{p_G}\cdot\sum_{(1-\delta_{\max})C\leqslant t\leqslant C}\left(\binom{G-1}{t-1}\binom{n-G}{C-t}\left(\binom{n-1}{C-1}\right)^{-1}\right).
\end{aligned}$$

Now we can continue with inequalities:

$$\begin{aligned}
B_1' &\leqslant \frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\mathbb{E}_k\left[\left\|\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2\right]\mid[3]\right]\\
&\leqslant \frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\mathbb{E}_k\left[\mathbb{E}_Q\left[\left\|\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2\right]\right]\mid[3]\right]\\
&\leqslant \frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\mathbb{E}_k\left[\mathbb{E}_Q\left[\left\|\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2\right]\right]\mid[3]\right]\\
&+\frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2\right]\mid[3]\right].
\end{aligned}$$

Using variance decomposition we have

$$\begin{aligned}
B_1' &\leqslant \frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\mathbb{E}_k\left[\mathbb{E}_Q\left[\left\|\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right\|^2\right]\right]-\sum_{i\in\mathcal{G}}\left\|\Delta_i\left(x^{k+1},x^k\right)\right\|^2\mid[3]\right]\\
&+\frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2\right]\mid[3]\right].
\end{aligned}$$

Applying Definition of unbiased compressor we have

$$\begin{aligned}
B_1' &\leqslant \frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}(1+\omega)\mathbb{E}_k\left\|\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right\|^2-\sum_{i\in\mathcal{G}}\left\|\Delta_i\left(x^{k+1},x^k\right)\right\|^2\mid[3]\right]\\
&+\frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2\mid[3]\right]\\
&\leqslant \frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}(1+\omega)\mathbb{E}_k\left\|\widehat{\Delta}_i\left(x^{k+1},x^k\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2\right]\\
&+\frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}(1+\omega)\mathbb{E}_k\left\|\Delta_i\left(x^{k+1},x^k\right)\right\|^2-\sum_{i\in\mathcal{G}}\mathbb{E}_k\left\|\Delta_i\left(x^{k+1},x^k\right)\right\|^2\mid[3]\right]\\
&+\frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2\mid[3]\right].
\end{aligned}$$

Now we combine terms and have

$$\begin{aligned}
B_1' &\leqslant \frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}(1+\omega)\mathbb{E}\left[\sum_{i\in\mathcal{G}}\mathbb{E}_k\left[\left\|\widehat{\Delta}_i\left(x^{k+1},x^k\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2\right]\mid[3]\right]\\
&+\frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\omega\mathbb{E}\left[\sum_{i\in\mathcal{G}}\left\|\Delta_i\left(x^{k+1},x^k\right)\right\|^2\mid[3]\right]\\
&+\frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2\mid[3]\right]\\
&= \frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}(1+\omega)\mathbb{E}\left[\sum_{i\in\mathcal{G}}\mathbb{E}_k\left[\left\|\widehat{\Delta}_i\left(x^{k+1},x^k\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2\right]\mid[3]\right]\\
&+\frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\omega\mathbb{E}\left[\sum_{i\in\mathcal{G}}\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2+\left\|\Delta\left(x^{k+1},x^k\right)\right\|^2\mid[3]\right]\\
&+\frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\mathbb{E}\left[\sum_{i\in\mathcal{G}}\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2\mid[3]\right].
\end{aligned}$$

Rearranging terms leads to

$$B_1' \leqslant \frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}(1+\omega)\mathbb{E}\left[\sum_{i\in\mathcal{G}}\mathbb{E}_k\left[\left\|\widehat{\Delta}_i\left(x^{k+1},x^k\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2\right] \mid [3]\right]$$
$$+\frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}(\omega+1)\mathbb{E}\left[\sum_{i\in\mathcal{G}}\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]$$
$$+\frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\omega\mathbb{E}\left[\sum_{i\in\mathcal{G}}\left\|\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right].$$

Now we apply Assumptions 4, 5, 6:

$$B_1' \leqslant \frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}(1+\omega)\mathbb{E}\left[G\frac{\mathcal{L}_\pm^2}{b}\|x^{k+1}-x^k\|^2\right]$$
$$+\frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}(\omega+1)\mathbb{E}\left[GL_\pm^2\|x^{k+1}-x^k\|^2\right]+\frac{\mathcal{P}_{\mathcal{G}_C^k}}{C(1-\delta_{\max})}\omega\mathbb{E}\left[GL^2\left\|x^{k+1}-x^k\right\|^2\right].$$

Finally we have

$$B_1' \leqslant \frac{\mathcal{P}_{\mathcal{G}_C^k}\cdot G}{C(1-\delta_{\max})}\left(\omega L^2+(\omega+1)L_\pm^2+\frac{(\omega+1)\mathcal{L}_\pm^2}{b}\right)\mathbb{E}\left[\|x^{k+1}-x^k\|^2\right].$$

Let us plug obtained results:

$$B_1 \leqslant \mathbb{E}\left[\left\|g^k-\nabla f(x^k)\right\|^2\right]$$
$$+\frac{\mathcal{P}_{\mathcal{G}_C^k}\cdot G}{C(1-\delta_{\max})}\left(\omega L^2+(\omega+1)L_\pm^2+\frac{(\omega+1)\mathcal{L}_\pm^2}{b}\right)\mathbb{E}\left[\|x^{k+1}-x^k\|^2\right].$$

Also we have

$$A_1 = \mathbb{E}\left[\left\|\overline{g}^{k+1}-\nabla f(x^{k+1})\right\|^2\right]$$
$$\leqslant (1-p)p_G B_1+(1-p)(1-p_G)\mathbb{E}\left[\left\|g^k-\nabla f(x^k)\right\|^2\right]$$
$$\leqslant (1-p)p_G\mathbb{E}\left[\left\|g^k-\nabla f(x^k)\right\|^2\right]$$
$$+(1-p)p_G\frac{\mathcal{P}_{\mathcal{G}_C^k}\cdot G}{C(1-\delta_{\max})}\left(\omega L^2+(\omega+1)L_\pm^2+\frac{(\omega+1)\mathcal{L}_\pm^2}{b}\right)\mathbb{E}\left[\|x^{k+1}-x^k\|^2\right]$$
$$+(1-p)(1-p_G)\mathbb{E}\left[\left\|g^k-\nabla f(x^k)\right\|^2\right].$$

Finally we get

$$A_1 \leqslant (1-p)\mathbb{E}\left[\left\|g^k-\nabla f(x^k)\right\|^2\right]$$
$$+(1-p)p_G\frac{\mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}\left(\omega L^2+(\omega+1)L_\pm^2+\frac{(\omega+1)\mathcal{L}_\pm^2}{b}\right)\mathbb{E}\left[\|x^{k+1}-x^k\|^2\right].$$

□

Lemma H.2. *Let Assumptions 4, 5, 6, 2 hold and the Compression Operator satisfy Definition 2.2. Also let us introduce the notation*

$$\texttt{ARAgg}_Q^{k+1} = \texttt{ARAgg}\left(\text{clip}_{\lambda_{k+1}}\left(\mathcal{Q}\left(\widehat{\Delta}_1(x^{k+1},x^k)\right)\right),\ldots,\text{clip}_{\lambda_{k+1}}\left(\mathcal{Q}\left(\widehat{\Delta}_C(x^{k+1},x^k)\right)\right)\right).$$

Then for all $k \geqslant 0$ the iterates produced by Byz-VR-MARINA-PP *(Algorithm 1) satisfy*

$$T_2 = \mathbb{E}\left[\mathbb{E}_k\left[\left\|\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k}\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right)-\texttt{ARAgg}_Q^{k+1}\right\|^2 \mid [3]\right]\right]$$
$$\leqslant 4\mathcal{P}_{\mathcal{G}_C^k}\left((1+\omega)\frac{\mathcal{L}_\pm^2}{b}+(\omega+1)L_\pm^2+\omega L^2\right)c\delta_{\max}\mathbb{E}\left[\|x^{k+1}-x^k\|^2\right],$$

where $\mathcal{P}_{\mathcal{G}_C^k} = \text{Prob}\left\{i \in \mathcal{G}_C^k \mid G_C^k \geqslant (1-\delta_{\max})\, C\right\}$.

Proof. Let us consider second term, since

$$
\begin{aligned}
T_2 &= \mathbb{E}\left[\mathbb{E}_k\left[\left\|\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k}\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right) - \texttt{ARAgg}_Q^{k+1}\right\|^2 \mid [3]\right]\right] \\
&\leqslant \mathbb{E}\left[\frac{c\delta_{\max}}{D_2}\sum_{\substack{i,l\in\mathcal{G}_C^k\\ i\neq l}}\mathbb{E}_k\left[\left\|\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right) - \text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_l\left(x^{k+1},x^k\right)\right)\right)\right\|^2 \mid [3]\right]\right],
\end{aligned}
$$

where $D_2 = G_C^k(G_C^k - 1)$.

Using $\lambda_{k+1} = D_Q \max_{i,j} L_{i,j}\|x^{k+1} - x^k\|$ we can guarantee that clipping operator becomes identical since we have

$$
\begin{aligned}
\left\|\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right\| &\leqslant D_Q\left\|\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right\| \\
&\leqslant D_Q\left\|\frac{1}{b}\sum_{j\in m}\nabla f_{i,j}(x^{k+1}) - \nabla f_{i,j}(x^k)\right\| \\
&\leqslant D_Q\frac{1}{b}\sum_{j\in m}\left\|\nabla f_{i,j}(x^{k+1}) - \nabla f_{i,j}(x^k)\right\| \\
&\leqslant D_Q\max_j L_{i,j}\left\|x^{k+1} - x^k\right\|
\end{aligned}
$$

Let us consider pair-wise differences:

$$
\begin{aligned}
T_2'(i,l) &= \mathbb{E}_k\left[\left\|\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right) - \text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_l\left(x^{k+1},x^k\right)\right)\right)\right\|^2 \mid [3]\right] \\
&\leqslant \mathbb{E}_k\left[\left\|\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right) - \Delta_i\left(x^{k+1},x^k\right) + \Delta_l\left(x^{k+1},x^k\right) - \text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_l\left(x^{k+1},x^k\right)\right)\right)\right\|^2 \mid [3]\right] \\
&+ \mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right) - \Delta_l\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right] \\
&\overset{(7)}{\leqslant} 2\mathbb{E}_k\left[\left\|\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right) - \Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right] \\
&+ 2\mathbb{E}_k\left[\left\|\Delta_l\left(x^{k+1},x^k\right) - \text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_l\left(x^{k+1},x^k\right)\right)\right)\right\|^2 \mid [3]\right] \\
&+ \mathbb{E}_k\left[\left\|\Delta_l\left(x^{k+1},x^k\right) - \Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]] \\
&\overset{(7)}{\leqslant} 2\mathbb{E}_k\left[\left\|\text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right) - \Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right] \\
&+ 2\mathbb{E}_k\left[\left\|\Delta_l\left(x^{k+1},x^k\right) - \text{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_l\left(x^{k+1},x^k\right)\right)\right)\right\|^2 \mid [3]\right] \\
&+ 2\mathbb{E}_k\left[\left\|\Delta_l\left(x^{k+1},x^k\right) - \Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right] \\
&+ 2\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right) - \Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right].
\end{aligned}
$$

Now we can combine all parts together:

$$\begin{aligned}
\widehat{T}_2 &= \mathbb{E}\left[\frac{1}{G_C^k(G_C^k-1)}\sum_{\substack{i,l\in\mathcal{G}_C^k\\ i\neq l}} T_2'(i,l)\right] \\
&\leqslant \mathbb{E}\left[\frac{1}{D_2}\sum_{\substack{i,l\in\mathcal{G}_C^k\\ i\neq l}} 2\mathbb{E}_k\left[\left\|\mathrm{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
&+ \mathbb{E}\left[\frac{1}{D_2}\sum_{\substack{i,l\in\mathcal{G}_C^k\\ i\neq l}} 2\mathbb{E}_k\left[\left\|\Delta_l\left(x^{k+1},x^k\right)-\mathrm{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_l\left(x^{k+1},x^k\right)\right)\right)\right\|^2 \mid [3]\right]\right] \\
&+ \mathbb{E}\left[\frac{1}{D_2}\sum_{\substack{i,l\in\mathcal{G}_C^k\\ i\neq l}} 2\mathbb{E}_k\left[\left\|\Delta_l\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
&+ \mathbb{E}\left[\frac{1}{D_2}\sum_{\substack{i,l\in\mathcal{G}_C^k\\ i\neq l}} 2\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right].
\end{aligned}$$

Combining terms together we have

$$\begin{aligned}
\widehat{T}_2 &\leqslant \mathbb{E}\left[\frac{1}{D_2}\sum_{\substack{i,l\in\mathcal{G}_C^k\\ i\neq l}} 4\mathbb{E}_k\left[\left\|\mathrm{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right)-\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
&+ \mathbb{E}\left[\frac{1}{D_2}\sum_{\substack{i,l\in\mathcal{G}_C^k\\ i\neq l}} 4\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right].
\end{aligned}$$

Using variance decomposition we get

$$\begin{aligned}
\widehat{T}_2 &\leqslant \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 4\mathbb{E}_k\left[\left\|\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right\|^2 \mid [3]\right]\right] \\
&- \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 4\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
&+ \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 4\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)-\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right].
\end{aligned}$$

Using properties of unbiased compressors we have

$$\begin{aligned}
\widehat{T}_2 \leqslant\ & \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 4(1+\omega)\mathbb{E}_k\left[\left\|\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
& - \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 4\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
& + \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 4\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right) - \Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
\leqslant\ & \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 4(1+\omega)\mathbb{E}_k\left[\left\|\widehat{\Delta}_i\left(x^{k+1},x^k\right) - \Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
& + \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 4(1+\omega)\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
& - \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 4\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
& + \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 4\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right) - \Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right].
\end{aligned}$$

Let us simplify the inequality:

$$\begin{aligned}
\widehat{T}_2 \leqslant\ & \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 4(1+\omega)\mathbb{E}_k\left[\left\|\widehat{\Delta}_i\left(x^{k+1},x^k\right) - \Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
& + \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 4\omega\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
& + \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 4\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right) - \Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right].
\end{aligned}$$

Using decomposition we have

$$\begin{aligned}
\widehat{T}_2 \leqslant\ & \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 4(1+\omega)\mathbb{E}_k\left[\left\|\widehat{\Delta}_i\left(x^{k+1},x^k\right) - \Delta_i\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
& + \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 4\omega\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right) - \Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
& + \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 4\mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1},x^k\right) - \Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right] \\
& + \mathbb{E}\left[\frac{1}{G_C^k}\sum_{i\in\mathcal{G}_C^k} 4\omega\mathbb{E}_k\left[\left\|\Delta\left(x^{k+1},x^k\right)\right\|^2 \mid [3]\right]\right].
\end{aligned}$$

Using similar argument in previous lemma we obtain

$$\begin{aligned}
\widehat{T}_2 \leqslant & \mathbb{E}\left[\frac{\mathcal{P}_{\mathcal{G}_C^k}}{G} \sum_{i \in \mathcal{G}} 4(1+\omega) \mathbb{E}_k\left[\left\|\widehat{\Delta}_i\left(x^{k+1}, x^k\right)-\Delta_i\left(x^{k+1}, x^k\right)\right\|^2 \mid [3]\right]\right] \\
& +\mathbb{E}\left[\frac{\mathcal{P}_{\mathcal{G}_C^k}}{G} \sum_{i \in \mathcal{G}} 4 \omega \mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1}, x^k\right)-\Delta\left(x^{k+1}, x^k\right)\right\|^2 \mid [3]\right]\right] \\
& +\mathbb{E}\left[\frac{\mathcal{P}_{\mathcal{G}_C^k}}{G} \sum_{i \in \mathcal{G}} 4 \mathbb{E}_k\left[\left\|\Delta_i\left(x^{k+1}, x^k\right)-\Delta\left(x^{k+1}, x^k\right)\right\|^2 \mid [3]\right]\right] \\
& +\mathbb{E}\left[\frac{\mathcal{P}_{\mathcal{G}_C^k}}{G} \sum_{i \in \mathcal{G}} 4 \omega \mathbb{E}_k\left[\left\|\Delta\left(x^{k+1}, x^k\right)\right\|^2 \mid [3]\right]\right].
\end{aligned}$$

Using Assumptions 4, 5, 6:

$$\begin{aligned}
\widehat{T}_2 \leqslant & \mathbb{E}\left[4(1+\omega) \mathcal{P}_{\mathcal{G}_C^k} \frac{\mathcal{L}_{\pm}^2}{b}\|x^{k+1}-x^k\|^2\right] \\
& +\mathbb{E}\left[4(\omega+1) \mathcal{P}_{\mathcal{G}_C^k} \omega L_{\pm}^2\|x^{k+1}-x^k\|^2\right] \\
& +\mathbb{E}\left[4 \mathcal{P}_{\mathcal{G}_C^k} \omega L^2\|x^{k+1}-x^k\|^2\right].
\end{aligned}$$

Finally, we obtain

$$\begin{aligned}
T_2 & =\mathbb{E}\left[\mathbb{E}_k\left[\left\|\frac{1}{G_C^k} \sum_{i \in \mathcal{G}_C^k} \operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1}, x^k\right)\right)\right)-\texttt{ARAgg}_Q^{k+1}\right\|^2 \mid [3]\right]\right] \\
& \leqslant 4 \mathcal{P}_{\mathcal{G}_C^k}\left((1+\omega) \frac{\mathcal{L}_{\pm}^2}{b}+(\omega+1) L_{\pm}^2+\omega L^2\right) c \delta_{\max } \mathbb{E}\left[\|x^{k+1}-x^k\|^2\right].
\end{aligned}$$

□

Lemma H.3. *Let Assumptions 1, 4, 5, 6, 9, 2 hold and Compression Operator satisfy Definition 2.2. We set $\lambda_{k+1}=D_Q \max_{i,j} L_{i,j}$. Also let us introduce the notation*

$$\texttt{ARAgg}_Q^{k+1}=\texttt{ARAgg}\left(\operatorname{clip}_{\lambda_{k+1}}\left(\mathcal{Q}\left(\widehat{\Delta}_1(x^{k+1}, x^k)\right)\right), \ldots, \operatorname{clip}_{\lambda_{k+1}}\left(\mathcal{Q}\left(\widehat{\Delta}_C(x^{k+1}, x^k)\right)\right)\right).$$

Then for all $k \geqslant 0$ the iterates produced by Byz-VR-MARINA-PP *(Algorithm 1) satisfy*

$$\begin{aligned}
\mathbb{E}\left[\left\|g^{k+1}-\nabla f\left(x^{k+1}\right)\right\|^2\right] \leqslant & \left(1-\frac{p}{2}\right) \mathbb{E}\left[\left\|g^k-\nabla f\left(x^k\right)\right\|^2\right] \\
& +\widehat{B} \mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right]+\widehat{C} \zeta^2+\frac{p A}{4}\|x^{k+1}-x^k\|^2,
\end{aligned}$$

$$\begin{aligned}
A= & \frac{4}{p}\left(\frac{80}{p} \frac{p_G \mathcal{P}_{\mathcal{G}_C^k}(1-\delta) n}{C(1-\delta_{\max })} \omega+24 \frac{\delta_{\max } \mathcal{P}_{\mathcal{G}_{\widehat{C}}^k} c \delta_{\max }}{(1-\delta_{\max })} B+\frac{4}{p}(1-p_G)+\frac{160}{p} p_G \mathcal{P}_{\mathcal{G}_C^k} c \delta_{\max } \omega\right) L^2 \\
& +\frac{4}{p}\left(\frac{8}{p} \frac{p_G \mathcal{P}_{\mathcal{G}_C^k}(1-\delta) n}{C(1-\delta_{\max })}(10 \omega+1)+\frac{16}{p} p_G \mathcal{P}_{\mathcal{G}_C^k} c \delta_{\max }(10 \omega+1)\right) L_{\pm}^2 \\
& +\frac{4}{p}\left(\frac{160}{p} p_G \mathcal{P}_{\mathcal{G}_C^k}(1+\omega) c \delta_{\max }+\frac{80}{p} p_G \mathcal{P}_{\mathcal{G}_C^k}(1+\omega) \frac{(1-\delta) n}{C(1-\delta_{\max })}\right) \frac{\mathcal{L}_{\pm}^2}{b} \\
& +\frac{4}{p}\left(\frac{4}{p}(1-p_G) F_{\mathcal{A}}^2\left(\max_{i \in \mathcal{G}} L_i\right)^2\right),
\end{aligned}$$

$$\widehat{B}=2 \frac{\delta_{\max } \mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}}{(1-\delta_{\max })} B(12 c \delta_{\max }+1) \quad \widehat{C}=2 \frac{\delta_{\max } \mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}}{(1-\delta_{\max })} B(6 c \delta_{\max }+1),$$

and where $p_G=\operatorname{Prob}\left\{G_C^k \geqslant(1-\delta_{\max }) C\right\}$ and $\mathcal{P}_{\mathcal{G}_C^k}=\operatorname{Prob}\left\{i \in \mathcal{G}_C^k \mid G_C^k \geqslant(1-\delta_{\max }) C\right\}$.

Proof. Let us combine bounds for A_1 and A_2 together:

$$
\begin{aligned}
A_0 &= \mathbb{E}\left[\left\|g^{k+1} - \nabla f\left(x^{k+1}\right)\right\|^2\right] \\
&\leqslant \left(1+\frac{p}{2}\right)\mathbb{E}\left[\left\|\overline{g}^{k+1} - \nabla f\left(x^{k+1}\right)\right\|^2\right] + \left(1+\frac{2}{p}\right)\mathbb{E}\left[\left\|g^{k+1} - \overline{g}^{k+1}\right\|^2\right] \\
&\leqslant \left(1+\frac{p}{2}\right)A_1 + \left(1+\frac{2}{p}\right)A_2 \\
&\leqslant \left(1+\frac{p}{2}\right)(1-p)\mathbb{E}\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \\
&+ \left(1+\frac{p}{2}\right)(1-p)p_G\frac{\mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}\left(\omega L^2 + (\omega+1)L_\pm^2 + \frac{(\omega+1)\mathcal{L}_\pm^2}{b}\right)\mathbb{E}\left[\|x^{k+1}-x^k\|^2\right]. \\
&+ \left(1+\frac{2}{p}\right)p\mathbb{E}\left[\mathbb{E}_k\left[\left\|\texttt{ARAgg}\left(\nabla f_1(x^{k+1}),\ldots,\nabla f_n(x^{k+1})\right) - \nabla f(x^{k+1})\right\|^2\right] \mid [1]\right] \\
&+ \left(1+\frac{2}{p}\right)(1-p)p_G\mathbb{E}\left[\mathbb{E}_k\left[\left\|\frac{1}{G_c^k}\sum_{i\in\mathcal{G}_C^k}\operatorname{clip}_\lambda\left(\mathcal{Q}\left(\widehat{\Delta}_i\left(x^{k+1},x^k\right)\right)\right) - \texttt{ARAgg}_Q^{k+1}\right\|^2 \mid [3]\right]\right] \\
&+ \left(1+\frac{2}{p}\right)(1-p)(1-p_G)\mathbb{E}\left[\mathbb{E}_k\left[\left\|\nabla f(x^{k+1}) - \nabla f(x^k) - \texttt{ARAgg}_Q^{k+1}\right\|^2 \mid [2]\right]\right].
\end{aligned}
$$

Using Lemmas H.2 and H.1 and previous lemmas from General Analysis we have

$$
\begin{aligned}
A_0 &= \mathbb{E}\left[\left\|g^{k+1} - \nabla f\left(x^{k+1}\right)\right\|^2\right] \\
&\leqslant \left(1-\frac{p}{2}\right)\mathbb{E}\left[\left\|g^k - \nabla f\left(x^k\right)\right\|^2\right] \\
&+ \left(1-\frac{p}{2}\right)p_G\frac{\mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}\left(\omega L^2 + (\omega+1)L_\pm^2 + \frac{(\omega+1)\mathcal{L}_\pm^2}{b}\right)\mathbb{E}\left[\|x^{k+1}-x^k\|^2\right] \\
&+ (p+2)\left(8\frac{\delta n\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}c\delta_{\max}}{(1-\delta_{\max})\frac{\delta n}{\delta_{\max}}}B\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] + 8\frac{\delta n\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}c\delta_{\max}}{(1-\delta_{\max})\frac{\delta n}{\delta_{\max}}}BL^2\mathbb{E}\left[\left\|x^{k+1}-x^k\right\|^2\right] + 4\frac{\delta n\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}c\delta_{\max}}{(1-\delta_{\max})\frac{\delta n}{\delta_{\max}}}\zeta^2.\right) \\
&+ \frac{2}{p}p_G\mathbb{E}\left[4(1+\omega)\mathcal{P}_{\mathcal{G}_C^k}\frac{\mathcal{L}_\pm^2}{b}c\delta_{\max}\|x^{k+1}-x^k\|^2\right] \\
&+ \frac{2}{p}p_G\mathbb{E}\left[4(\omega+1)\mathcal{P}_{\mathcal{G}_C^k}L_\pm^2c\delta_{\max}\|x^{k+1}-x^k\|^2\right] \\
&+ \frac{2}{p}p_G\mathbb{E}\left[4\mathcal{P}_{\mathcal{G}_C^k}\omega L^2c\delta_{\max}\|x^{k+1}-x^k\|^2\right] + \frac{2}{p}(1-p_G)2(L^2 + F_{\mathcal{A}}^2\alpha_{\lambda_{k+1}}^2)\mathbb{E}\left[\left\|x^{k+1}-x^k\right\|^2\right].
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
\mathbb{E}\left[\left\|g^{k+1} - \nabla f\left(x^{k+1}\right)\right\|^2\right] &\leqslant \left(1-\frac{p}{2}\right)\mathbb{E}\left[\left\|g^k - \nabla f\left(x^k\right)\right\|^2\right] \\
&+ \widehat{B}\mathbb{E}\left[\left\|\nabla f\left(x^k\right)\right\|^2\right] + \widehat{C}\zeta^2 + \frac{pA}{4}\|x^{k+1}-x^k\|^2,
\end{aligned}
$$

where

$$
\begin{aligned}
A &= \frac{2}{p}\left(\frac{p_G\mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}\omega + 24c\delta B + \frac{4}{p}(1-p_G) + \frac{8}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}c\delta_{\max}\omega\right)L^2 \\
&+ \frac{2}{p}\left(\frac{p_G\mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}(\omega+1) + \frac{8}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}c\delta_{\max}(\omega+1)\right)L_\pm^2 \\
&+ \frac{2}{p}\left(p_G\mathcal{P}_{\mathcal{G}_C^k}(1+\omega)c\delta_{\max} + \frac{8}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}(1+\omega)\frac{(1-\delta)n}{C(1-\delta_{\max})}\right)\frac{\mathcal{L}_\pm^2}{b} \\
&+ \frac{2}{p}\left(\frac{4}{p}(1-p_G)F_{\mathcal{A}}^2\left(D_Q\max_{i,j}L_{i,j}\right)^2\right).
\end{aligned}
$$

□

Theorem H.1. *Let Assumptions 1, 4, 5, 6, 9, 2 hold. Setting* $\lambda_{k+1} = \max_{i,j} L_{i,j} \left\|x^{k+1} - x^k\right\|$. *Assume that*

$$0 < \gamma \leqslant \frac{1}{L + \sqrt{A}}, \quad 4\widehat{B} < p,$$

where where

$$\begin{aligned}
A = {} & \frac{4}{p}\left(\frac{80}{p}\frac{p_G\mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}\omega + 24\frac{\delta_{\max}\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}c\delta_{\max}}{(1-\delta_{\max})}B + \frac{4}{p}(1-p_G) + \frac{160}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}c\delta_{\max}\omega\right)L^2 \\
& + \frac{4}{p}\left(\frac{8}{p}\frac{p_G\mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}(10\omega+1) + \frac{16}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}c\delta_{\max}(10\omega+1)\right)L_{\pm}^2 \\
& + \frac{4}{p}\left(\frac{160}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}(1+\omega)c\delta_{\max} + \frac{80}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}(1+\omega)\frac{(1-\delta)n}{C(1-\delta_{\max})}\right)\frac{\mathcal{L}_{\pm}^2}{b} \\
& + \frac{4}{p}\left(\frac{4}{p}(1-p_G)F_{\mathcal{A}}^2\left(\max_{i\in\mathcal{G}} L_i\right)^2\right),
\end{aligned}$$

$$\widehat{B} = 2\frac{\delta_{\max}\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}}{(1-\delta_{\max})}B(12c\delta_{\max}+1) \quad \widehat{C} = 2\frac{\delta_{\max}\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}}{(1-\delta_{\max})}B(6c\delta_{\max}+1),$$

and

$$\begin{aligned}
\mathcal{P}_{\mathcal{G}_C^k} &= \frac{C}{np_G}\cdot\sum_{(1-\delta_{\max})C\leqslant t\leqslant C}\left(\binom{G-1}{t-1}\binom{n-G}{C-t}\left(\binom{n}{C}\right)^{-1}\right), \\
p_G &= \mathbb{P}\left\{G_C^k \geqslant (1-\delta_{\max})\,C\right\} \\
&= \sum_{\lceil(1-\delta_{\max})C\rceil\leqslant t\leqslant C}\left(\binom{G}{t}\binom{n-G}{C-t}\binom{n}{C}^{-1}\right),
\end{aligned}$$

Then for all $K \geqslant 0$ *the iterates produced by Byz-VR-MARINA (Algorithm 1) satisfy*

$$\mathbb{E}\left[\left\|\nabla f\left(\widehat{x}^K\right)\right\|^2\right] \leqslant \frac{2\Phi^0}{\gamma\left(1-\frac{4\widehat{B}}{p}\right)(K+1)} + \frac{2\widehat{C}\zeta^2}{p-4\widehat{B}},$$

where $\widehat{x}^K$ *is choosen uniformly at random from* $x^0, x^1, \ldots, x^K$, *and* $\Phi^0 = f\left(x^0\right) - f^* + \frac{\gamma}{p}\left\|g^0 - \nabla f\left(x^0\right)\right\|^2$.

Proof. The proof is analogous to proof of Theorem G.1. □

Theorem H.2. *Let Assumptions 1, 2, 4, 5, 6, 9, 8 hold. Setting* $\lambda_{k+1} = \max_{i,j} L_{i,j} \left\|x^{k+1} - x^k\right\|$. *Assume that*

$$0 < \gamma \leqslant \min\left\{\frac{1}{L + \sqrt{2A}}\right\}, \quad 8\widehat{B} < p,$$

where where

$$\begin{aligned}
A = {} & \frac{4}{p}\left(\frac{80}{p}\frac{p_G\mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}\omega + 24\frac{\delta_{\max}\mathcal{P}_{\mathcal{G}_{\widehat{C}}^k}c\delta_{\max}}{(1-\delta_{\max})}B + \frac{4}{p}(1-p_G) + \frac{160}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}c\delta_{\max}\omega\right)L^2 \\
& + \frac{4}{p}\left(\frac{8}{p}\frac{p_G\mathcal{P}_{\mathcal{G}_C^k}(1-\delta)n}{C(1-\delta_{\max})}(10\omega+1) + \frac{16}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}c\delta_{\max}(10\omega+1)\right)L_{\pm}^2 \\
& + \frac{4}{p}\left(\frac{160}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}(1+\omega)c\delta_{\max} + \frac{80}{p}p_G\mathcal{P}_{\mathcal{G}_C^k}(1+\omega)\frac{(1-\delta)n}{C(1-\delta_{\max})}\right)\frac{\mathcal{L}_{\pm}^2}{b} \\
& + \frac{4}{p}\left(\frac{4}{p}(1-p_G)F_{\mathcal{A}}^2\left(\max_{i\in\mathcal{G}} L_i\right)^2\right),
\end{aligned}$$

$$\widehat{B} = 2\frac{\delta_{\max}\mathcal{P}_{\mathcal{G}^k_{\widehat{C}}}}{(1-\delta_{\max})}B(12c\delta_{\max}+1) \quad \widehat{C} = 2\frac{\delta_{\max}\mathcal{P}_{\mathcal{G}^k_{\widehat{C}}}}{(1-\delta_{\max})}B(6c\delta_{\max}+1),$$

and where $p_G = \text{Prob}\left\{G^k_C \geqslant (1-\delta_{\max})C\right\}$ *and* $\mathcal{P}_{\mathcal{G}^k_C} = \text{Prob}\left\{i \in \mathcal{G}^k_C \mid G^k_C \geqslant (1-\delta_{\max})\, C\right\}$.
Then for all $K \geqslant 0$ *the iterates produced by Byz-VR-MARINA (Algorithm 1) satisfy*

$$\mathbb{E}\left[f\left(x^K\right) - f\left(x^*\right)\right] \leqslant (1-\rho)^K \Phi^0 + \frac{2\widehat{C}\zeta^2}{p\rho},$$

where $\rho = \min\left[\gamma\mu\left(1 - \frac{8\widehat{B}}{p}\right), \frac{p}{4}\right]$ *and* $\Phi^0 = f\left(x^0\right) - f^* + \frac{2\gamma}{p}\left\|g^0 - \nabla f\left(x^0\right)\right\|^2$.

Proof. The proof is analogous to proof of Theorem G.2. □