
Byzantine-Robust and Differentially Private Federated Optimization under Weaker Assumptions

Rustem Islamov¹ Grigory Malinovsky² Alexander Gaponov² Aurelien Lucchi¹ Peter Richtárik²
Eduard Gorbunov³

¹ University of Basel, Switzerland

² KAUST, Saudi Arabia

³ MBZUAI, UAE

Abstract

Federated Learning (FL) enables heterogeneous clients to collaboratively train a shared model without centralizing their raw data, offering an inherent level of privacy. However, gradients and model updates can still leak sensitive information, while malicious clients or other adversarial participants may mount attacks such as Byzantine manipulation. These vulnerabilities highlight the need to address differential privacy (DP) and Byzantine robustness within a unified framework. Existing approaches, however, often rely on unrealistic assumptions such as bounded gradients, require auxiliary server-side datasets, or fail to provide convergence guarantees. We address these limitations by proposing Byz-Clip21-SGD2M, a new algorithm that integrates robust aggregation with double momentum and carefully designed clipping. We prove high-probability convergence guarantees under standard L -smoothness and σ -sub-Gaussian gradient noise assumptions, thereby relaxing conditions that dominate prior work. Our analysis recovers state-of-the-art convergence rates in the absence of adversaries and improves utility guarantees under Byzantine and DP settings. Empirical evaluations on CNN and MLP models trained on MNIST further validate the effectiveness of our approach.

1 INTRODUCTION

The rapid deployment of large-scale machine learning models has positioned Federated Learning (FL) [Konečný et al., 2016, McMahan et al., 2017a] as a central paradigm, enabling a collection of potentially heterogeneous clients, ranging from smartphones and sensors to data centers, to collaboratively train a shared model without transmitting their raw

data to a central server [Li et al., 2020, Yang et al., 2019]. While this setting offers an inherent degree of privacy, it is far from complete: gradients and model parameters can still leak sensitive information [Zhu et al., 2019, Geiping et al., 2020], and adversaries may exploit the distributed nature of FL to launch attacks such as membership inference, model stealing, or Byzantine manipulation [Shokri et al., 2017, Tramèr et al., 2016, Blanchard et al., 2017]. This dual vulnerability highlights the need to study differential privacy, to mitigate privacy leakage, and Byzantine robustness, to ensure resilience against arbitrary or malicious participants, within a unified framework. Both challenges share the common difficulty of optimizing with corrupted or unreliable updates, whether the corruption arises from deliberate noise injection for privacy or from adversarial manipulations. Moreover, addressing these issues is not only of algorithmic importance: modern AI systems must also comply with evolving regulatory requirements, such as GDPR and the EU AI Act for privacy [GDP, 2016, EUA, 2024], or the EU NIS2 Directive, Cyber Resilience Act, and NIST AI Risk Management Framework for security and robustness [NIS, 2022, Cyb, 2024, Tabassi et al., 2023].

In this context, we propose to study Byzantine robustness and differential privacy jointly, with the goal of establishing a principled foundation for robust, privacy-preserving federated learning and providing a pathway toward trustworthy AI in practice. A central open challenge is to move beyond the strong assumptions that dominate the current literature. Most theoretical analyses in differential privacy are carried out under assumptions such as bounded gradients – an assumption that fails even for simple quadratic objectives [Li et al., 2022, Wang et al., 2023, Lowy et al., 2023] – or by focusing on the full-batch regime [Shulgin et al., 2025b]. Only recently have approaches emerged that achieve strong privacy guarantees while retaining favorable optimization performance under milder conditions [Islamov et al., 2025b]. Byzantine robustness has evolved largely as a separate line of work [Lyu et al., 2022], yet its analyses similarly depend on gradient boundedness to establish con-

vergence guarantees [Blanchard et al., 2017, Mhamdi et al., 2018]. Attempts to combine DP with Byzantine resilience remain even more limited: existing methods either assume access to an auxiliary server-side dataset to obtain a better gradient approximation [Xiang et al., 2023], rely on aggregation rules or modeling assumptions that are difficult to realize in this setting [Guerraoui et al., 2021b, Zhang and Hu, 2023], and impose restrictive bounded-gradient assumptions [Allouah et al., 2023b].

Motivated by the shortcomings of prior theoretical results, we arrive at the following guiding question:

Is it possible to develop an algorithm that provably converges in the presence of malicious clients while simultaneously providing strong differential privacy guarantees under standard assumptions?

Main Contributions. We provide an affirmative answer to the above question. In doing so, we make the following contributions:

- We introduce Byz-Clip21-SGD2M, a new algorithm that combines robust aggregation with double momentum and carefully designed clipping, achieving efficient performance under both Byzantine and differential privacy adversaries.
- We provide a high-probability convergence analysis of Byz-Clip21-SGD2M under L -smoothness and σ -sub-Gaussian gradient noise, thereby relaxing assumptions that have been used in prior work (see Table 1 for details). Our results improve utility guarantees by avoiding unrealistic assumptions, while recovering state-of-the-art convergence rates in the absence of Byzantine or DP adversaries.
- We complement our theoretical analysis with empirical validation, demonstrating the effectiveness of Byz-Clip21-SGD2M on CNN and MLP models trained on the MNIST dataset.

2 RELATED WORKS

Error Feedback. Biased compression in federated learning reduces communication by transmitting compressed local updates that are not unbiased estimates of the original signal [Ajalloeian and Stich, 2020, Beznosikov et al., 2023, Demidovich et al., 2023]. Common examples include Top- K sparsification [Alistarh et al., 2017, Wangni et al., 2018] and sign-based compression [Bernstein et al., 2018, Stich et al., 2018]. While these methods are efficient in practice, the introduced bias can accumulate and affect convergence, requiring Error Feedback (EF) [Seide et al., 2014] techniques to mitigate its impact. Earlier works primarily focused on the single-node setting or relied on restrictive assumptions, such as bounded gradients, bounded compression error, or gradient dissimilarity, to establish convergence

[Stich et al., 2018, Stich and Karimireddy, 2019, Koloskova et al., 2019]. Moreover, the convergence rates of EF degrade in the presence of client heterogeneity, and this dependence is intrinsic rather than a proof artifact [Gorbunov et al., 2020]. To overcome these limitations, [Richtárik et al., 2021] proposed EF21, a variant whose guarantees do not depend on heterogeneity bounds. Nevertheless, EF21-SGD still requires increasingly large batch sizes to achieve a target accuracy [Fatkhullin et al., 2024]. Importantly, this limitation is not fundamental: recent work demonstrates that incorporating Heavy-Ball momentum removes the need for large batches [Fatkhullin et al., 2024]. Later, EF21 was extended to several practical setups [Fatkhullin et al., 2025, Makarenko et al., 2022]. In a parallel line of research, Gao et al. [2023, 2025b] introduced EControl-type Error Feedback methods with convergence guarantees extending to the convex setting. Error Feedback has also been studied in decentralized [Yau and Wai, 2022, Huang et al., 2023, Islamov et al., 2024a] setting, for non-smooth [Islamov et al., 2025a] and composite [Gao et al., 2025a] optimization problems, as well as in the context of second-order methods [Safaryan et al., 2021, Qian et al., 2021, Islamov et al., 2023].

Byzantine Robust Optimization. Naively averaging updates from clients in a distributed setting lacks robustness since even a single Byzantine client can destabilize training [Chen et al., 2017]. To address this, numerous alternatives to averaging have been proposed to enhance robustness against adversarial updates [Pillutla et al., 2022, Yin et al., 2018, Damaskinos et al., 2019, Karimireddy et al., 2021, Allouah et al., 2024a].

Nevertheless, permutation-invariant algorithms, i.e., the ones whose output is unchanged under shuffling of the stochastic gradients computed on different workers [Karimireddy et al., 2021, Definition B], alone cannot ensure convergence in the Byzantine regime [Karimireddy et al., 2021, Theorem II]. This limitation can be alleviated by incorporating update shuffling and averaging in random groups [Karimireddy et al., 2020], variance-reduction techniques [Wu et al., 2020, Gorbunov et al., 2022b], or preprocessing based on the averaging with nearest neighbors [Allouah et al., 2023a]. Beyond these, other approaches include concentration-based gradient filtering [Alistarh et al., 2018, Allen-Zhu et al., 2020], server-free protocols with random checks [Gorbunov et al., 2022a], redundant computations [Chen et al., 2018, Rajput et al., 2019], and reputation-based mechanisms [Rodriguez-Barroso et al., 2020, Regatti et al., 2020, Xu and Lyu, 2020]. More recent works further extend these ideas to partial participation [Malinovsky et al., 2024], decentralized training [He et al., 2022], and communication compression [Rammal et al., 2024]. In most of these cases, the theoretical analysis is carried out under the L -smoothness and ζ^2 -gradient dissimilarity¹ assumptions, which have become standard in Byzantine robustness

¹More precisely, the mentioned works rely on some variation

Table 1: Convergence guarantees of algorithms in the presence of Byzantine and/or DP adversaries. $\tilde{O}$ hides logarithmic and decaying-with- T terms. Notation: (s)CVX=(strongly) convex functions, nCVX=non-convex functions, DP=supports differential privacy, BR=supports Byzantine robustness, E=in-expectation analysis, P=high-probability analysis.

Method	Utility	DP	BR	Assumptions	Additional Comments	Type
SoteriaFL [Liu et al., 2022]	$\tilde{O}\left(\frac{\sqrt{d}}{\sqrt{G\varepsilon}}\right)$	✓	✗	M -bounded gradients	nCVX, local DP	E
α -Norm [Shulgin et al., 2025a]	$\tilde{O}\left(\frac{\sqrt{d}}{\sqrt{G\varepsilon}} + R\right)^{(a)}$	✓	✗	L -smoothness Full-batch gradients	nCVX, local DP	E
Clip21-SGD2M [Islamov et al., 2025b]	$\tilde{O}\left(\frac{\sqrt{d}}{\sqrt{G\varepsilon}}\right)^{(b)}$	✓	✗	L -smoothness σ -sub-Gaussian noise	nCVX, local DP	P
Byz-SGDM [Karimireddy et al., 2020]	$\tilde{O}(\delta_{\text{byz}}\zeta^2)^{(c)}$	✗	✓	L -smoothness Bounded variance	nCVX	E
Robust D-SHB [Allouah et al., 2023a]	$\tilde{O}(\delta_{\text{byz}}M^2)$	✗	✓	L -smoothness Full-batch gradients	nCVX CVX	E
Byz-VR-MARINA [Gorbunov et al., 2022b]	$\tilde{O}(\delta_{\text{byz}}\zeta^2)$	✗	✓	L -smoothness Local functions are finite-sums	nCVX	E
Safe-DSHB [Allouah et al., 2023b]	$\tilde{O}\left(\frac{d}{G\varepsilon^2} + \frac{\delta_{\text{byz}}}{\varepsilon^2} + \delta_{\text{byz}}M^2\right)$	✓	✓	M -bounded gradients	sCVX, local DP	E
DP-BREM [Gu et al., 2025]	$\tilde{O}\left(\delta_{\text{byz}}\zeta^2 + \delta_{\text{byz}}\frac{\sqrt{d}}{\sqrt{G\varepsilon}}\right)^{(d)}$	✓	✓	L -smoothness Momentum buffers are i.i.d	nCVX, central DP	E
P&S Learning [Xiang et al., 2023]	N/A ^(e)	✓	✓	L -smoothness $\nabla f_i(x) - \nabla f_i(x^*)$ are sub-Gaussian Access to external non-private dataset	sCVX, local DP	P
Byz-Clip21-SGD2M [Ours]	$\tilde{O}\left(R + R^{2/3}\right)^{(b,f)}$ $R = \frac{\sqrt{d}}{\sqrt{G\varepsilon}} + \frac{\sqrt{\delta_{\text{byz}}d}}{\varepsilon} + \sqrt{\delta_{\text{byz}}}\mathcal{N}$	✓	✓	L -smoothness σ -sub-Gaussian noise	nCVX, local DP	P

(a) Here $R := \max_{i \in \mathcal{G}} \|\nabla f_i(x^0) - g_i^0\|$.

(b) Derived under the assumption $\sqrt{d}/\sqrt{G\varepsilon} \geq 1$.

(c) ζ^2 denotes the constant in the gradient dissimilarity assumption.

(d) The provided rate contradicts the lower bound due to the unrealistic assumption that momentum buffers are i.i.d. distributed.

(e) The authors do not provide explicit bounds.

(f) In the general case, $\mathcal{N} = LF^0 + \zeta^2 + \sigma(\sqrt{LF^0} + \zeta)$, where $F^0 = f(x^0) - f^*$ (see Theorem 5.1). In the presence of Byzantine adversaries only, $\mathcal{N} = \sqrt{\delta_{\text{byz}}}\zeta^2$ (see Theorem 5.6).

literature.

Differentially Private Optimization. Differential privacy (DP) is typically achieved by clipping each client’s update and adding Gaussian noise, limiting the influence of any individual client [McMahan et al., 2017b]. In central DP, a trusted server adds noise before updating the global model, while in local DP, clients perturb updates before sending them, protecting privacy even from the server [Kasiviswanathan et al., 2011, Allouah et al., 2024b]. Local DP offers stronger privacy but reduces model utility, though this can be mitigated using secure shufflers or aggregators [Erlingsson et al., 2019, Bonawitz et al., 2017]. DP can also enable update compression without additional cost [Chaudhuri et al., 2022, Hegazy et al., 2023]. Private optimization methods, such as DP-SGD [Abadi et al., 2016], enforce DP by clipping gradients and adding noise scaled to the clipping sensitivity. However, most convergence analyses neglect the bias introduced by clipping. For smooth functions, guarantees typically assume either *bounded gradient norms* [Li et al., 2020, Zhang et al., 2020b, Murata and Suzuki, 2023, Wang et al., 2023, Lowy et al., 2023, Wang et al., 2024] or that *clipping is effectively inactive* [Noble et al., 2022, Zhang et al., 2024]. Recently, several works

have provided improved analyses of DP-Clip-SGD and its variants. In the single-node setting, generalized smoothness has also been considered without restrictive requirements for DP-Clip-SGD [Koloskova et al., 2023]. Next, Khah et al. [2025] also focus on the single-node setting and provide high-probability results under the heavy-tailed noise assumption and arbitrary clipping level for DP-Clip-SGD. Zhao et al. [2025] consider a projected SGD-type method utilizing DP mean estimation and provide its in-expectation convergence bounds. Compagnoni et al. [2026] study SDE approximation of DP-Clip-SGD in the single-node setup. In the FL setting, analysis under realistic conditions in the presence of differential privacy has been studied in Islamov et al. [2025b], Shulgin et al. [2025a,b].

Differentially Private and Byzantine Robust Methods.

Although substantial progress has been made separately in differential privacy and Byzantine robustness, their interaction in distributed learning is still not well understood. Recent efforts to address both aspects at once remain limited. A key limitation of existing theoretical work is the reliance on restrictive assumptions, often unrealistic and mainly introduced to simplify the analysis. For example, some studies focus on a single aggregation rule [Guerraoui et al., 2021b], while others assume a particular noise structure near the min-

of Assumption 3.5.

imizer or even access to a non-private external server-side dataset [Xiang et al., 2023]. Other analyses restrict attention to the strongly convex case [Gao et al., 2024, Allouah et al., 2023b] or impose functional similarity assumptions that do not hold in heterogeneous federated learning [Lan et al., 2025]. Further limitations include assuming i.i.d. momentum buffers [Gu et al., 2025] or bounded-gradient conditions [Guerraoui et al., 2021a, Gao et al., 2024, Allouah et al., 2023b]. Taken together, these limitations highlight the significant challenges in jointly addressing differential privacy and Byzantine robustness.

We summarize the existing works and the assumptions underlying their convergence analyses for establishing utility bounds in Table 1. Compared to prior work, our theoretical analysis relies on the weakest set of assumptions.

3 PRELIMINARIES

Problem Formulation. In distributed training with potentially Byzantine clients, the goal is to minimize the empirical loss

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

where $x \in \mathbb{R}^d$ denotes the model parameters. The set of regular clients is $\mathcal{G}$, with cardinality $G := |\mathcal{G}|$, while the total number of clients is n . The remaining clients, $\mathcal{B} := [n] \setminus \mathcal{G}$, are Byzantine: instead of following the prescribed learning algorithm, they may transmit arbitrary (possibly adversarial) vectors, and we assume they also observe the updates shared by the other clients². The number of Byzantine clients is $|\mathcal{B}|$ and it is bounded as $|\mathcal{B}| \leq \delta_{\text{byz}} n$, with $\delta_{\text{byz}} < \frac{1}{2}$ to ensure that Byzantine clients do not form a majority; otherwise, problem (1) becomes unsolvable [Pai et al., 2021]. We assume that δ_{byz} is known to the server. Each client $i \in [n]$ is associated with a local loss function f_i , defined over its own data.

Differential Privacy. Next, we will use the following classical definition of (ε, δ) -Differential Privacy $((\varepsilon, \delta)$ -DP), which introduces plausible deniability into the output of a learning algorithm.

Definition 3.1 ([Dwork et al., 2014]). A randomized method $\mathcal{M}: \mathcal{D} \rightarrow \mathbb{R}$ satisfies (ε, δ) -Differential Privacy $((\varepsilon, \delta)$ -DP for shortness) if for any two datasets $D, D' \in \mathcal{D}$ that differ in 1 sample and for any $S \subseteq \mathbb{R}$

$$\mathbb{P}(\mathcal{M}(D) \in S) \leq e^\varepsilon \mathbb{P}(\mathcal{M}(D') \in S) + \delta. \quad (2)$$

²This assumption is admittedly strong and often unrealistic in practice. Nevertheless, it is widely adopted in the Byzantine-robust learning literature, since methods that guarantee robustness against such powerful adversaries also remain robust in more practical settings, where attackers are typically weaker.

Decreasing ε and δ strengthens privacy by making it harder to identify the particular data point that differs between neighboring datasets.

Robust Aggregation. We follow the definition of Allouah et al. [2023a] that covers many practical aggregation rules introduced in prior work [Farhadkhani et al., 2022, Karimireddy et al., 2020], including coordinate-wise median [Yin et al., 2018] and geometric median [Pillutla et al., 2022].

Definition 3.2 ([Allouah et al., 2023a]). Let the ratio of Byzantine clients be $|\mathcal{B}|/n \leq \delta_{\text{byz}} < 1/2$ and $c \geq 0$ be a constant. The aggregation rule RAgg is said to be (δ_{byz}, c) -robust if for any vectors $\{x_1, \dots, x_n\} \subseteq \mathbb{R}^d$ and any subset $S \subset [n]$ of size $n - |\mathcal{B}|$ the output $\hat{x} = \text{RAgg}(x_1, \dots, x_n)$ satisfies

$$\|\hat{x} - \bar{x}\|^2 \leq \frac{c\delta_{\text{byz}}}{n - |\mathcal{B}|} \sum_{i \in S} \|x_i - \bar{x}\|^2 \quad (3)$$

where $\bar{x} = \frac{1}{n - |\mathcal{B}|} \sum_{i \in S} x_i$.

Under this definition, robust aggregators deliver a worst-case guarantee: for $S = \mathcal{G}$, the output $\hat{x}$ stays close to the mean $\bar{x}$ of regular clients in squared distance. The rule can be further strengthened by input mixing steps such as NNM [Allouah et al., 2023a]. Importantly, we require the worst-case property, i.e., (3) holds for every vector collection $\{x_i\}_{i=1}^n$ to enable our high-probability analysis. In contrast, prior work often studies stochastic aggregators under in-expectation guarantees [Karimireddy et al., 2021, 2020, Gorbunov et al., 2022b, Malinovsky et al., 2024].

Assumptions. In the convergence analysis of Byz-Clip21-SGD2M, we make use of two standard assumptions. The first describes the class of smooth functions, a standard assumption made in the non-convex optimization literature [Ghadimi and Lan, 2013, Carmon et al., 2020].

Assumption 3.3. For all $i \in \mathcal{G}$, f_i is L -smooth, i.e., for all $x, y \in \mathbb{R}^d$ we have

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|. \quad (4)$$

Additionally, we have $f^* := \inf_{x \in \mathbb{R}^d} f(x) > -\infty$.

To simplify the presentation, we consider the worst smoothness constant across clients. Next, we introduce the assumption that formalizes the stochasticity in local gradients.

Assumption 3.4. Each client $i \in \mathcal{G}$ has access to a σ -sub-Gaussian unbiased estimator $\nabla f_i(x, \xi)$ of a local gradient $\nabla f_i(x)$, i.e., there exists a constant $\sigma \geq 0$ such that for all $x \in \mathbb{R}^d$ we have

$$\mathbb{E}[\nabla f_i(x, \xi)] = \nabla f_i(x), \quad \mathbb{E}[\exp(\|\theta_i\|^2/\sigma^2)] \leq e, \quad (5)$$

where ξ denotes the source of stochasticity and $\theta_i := \nabla f_i(x, \xi) - \nabla f_i(x)$.

Sub-Gaussian noise assumption is standard when analyzing high-probability complexity of SGD-type algorithms [Ghadimi and Lan, 2012, Nemirovski et al., 2009, Liu et al., 2023].

Bounds on data heterogeneity are essential in the analysis of Byzantine-robust algorithms; without such assumptions, robustness is unachievable in fully arbitrary heterogeneous environments. A common approach is to assume a bounded average deviation of local gradients from the global gradient (gradient dissimilarity) [Wu et al., 2020, Gorbunov et al., 2022b, Allouah et al., 2023b], as formalized in Assumption 3.5. For our high-probability analysis, we instead require the maximum in Assumption 3.5a. However, we impose this condition only at initialization x^0 , whereas prior work typically assumes it holds uniformly over the entire space [Islamov et al., 2024b, Mishchenko et al., 2022].

Assumption 3.5. The local gradients $\{\nabla f_i(x)\}_{i \in \mathcal{G}}$ satisfy bounded ζ^2 -gradient dissimilarity for some $\zeta \geq 0$, i.e., for all $x \in \mathbb{R}^d$ we have

$$\frac{1}{G} \sum_{i \in \mathcal{G}} \|\nabla f_i(x) - \nabla f(x)\|^2 \leq \zeta^2. \quad (6)$$

Assumption 3.5a. The local gradients $\{\nabla f_i(x^0)\}_{i \in \mathcal{G}}$ satisfy bounded ζ^2 -gradient dissimilarity for some $\zeta \geq 0$, i.e., we have

$$\max_{i \in \mathcal{G}} \|\nabla f_i(x^0) - \nabla f(x^0)\|^2 \leq \zeta^2. \quad (7)$$

Remark 3.7. Our results also extend to a more general condition in which the right-hand side of (7) scales linearly with $\|\nabla f(x^0)\|^2$, namely, $\max_{i \in \mathcal{G}} \|\nabla f_i(x^0) - \nabla f(x^0)\|^2 \leq A_\zeta \|\nabla f(x^0)\|^2 + \zeta^2$. For clarity of presentation, the general result is provided in the supplementary materials, while we assume $A_\zeta = 0$ in the main text. Note that this assumption always holds, i.e., for any problem of the form (1), there exists $\zeta \geq 0$ such that inequality (7) is satisfied. However, as we explicitly show in our theoretical results, the value of ζ directly affects the convergence bounds since it is required to establish a bound Δ on the initial value of the Lyapunov function Φ^0 introduced in (9).

4 ALGORITHM DESIGN

We introduce the Byz-Clip21-SGD2M algorithm, designed to operate under both Byzantine failures and differential privacy noise. The main steps are outlined in Algorithm 1, followed by a detailed discussion highlighting the key design innovations and their rationale.

Momentum Mechanism. Byz-Clip21-SGD2M employs two heavy-ball momentum buffers, $\{v_i^t\}_{i \in \mathcal{G}}$ on the client side and $\{m_i^t\}_{i \in \mathcal{G}}$ on the server side, both essential for mitigating noise during training. On the client side, momentum

with parameter β smooths mini-batch gradient noise; without it, the method may fail to converge, as seen in algorithms such as EF21-SGD and Clip21-SGD, which provably fail to converge to arbitrary accuracy ε under mini-batch noise [Fatkhullin et al., 2024, Islamov et al., 2025b]. On the server side, momentum with parameter $\hat{\beta}$ counteracts the accumulation of DP noise in the buffers $\{m_i^t\}_{i \in \mathcal{G}}$, preventing it from destroying convergence due to fast accumulation of the DP noise in the updates.

Clipping. It is crucial to bound model updates before injecting DP noise. Without clipping, updates may have arbitrarily large norms, which would dominate the injected noise and invalidate the privacy guarantee. Clipping restricts each update to a controlled range, ensuring that the privacy guarantees remain meaningful and that the added noise achieves its intended effect. In Byz-Clip21-SGD2M, we employ the standard norm-clipping mechanism, which rescales updates relative to a threshold τ :

$$\text{clip}_\tau(x) = \begin{cases} \frac{\tau}{\|x\|}x, & \text{if } \|x\| > \tau, \\ x, & \text{if } \|x\| \leq \tau. \end{cases} \quad (8)$$

This procedure ensures a fixed sensitivity at each update, after which Gaussian noise is added to the clipped values to provide local (ε, δ) -DP guarantees. Equivalently, the same effect can be interpreted as adding noise directly to the averaged client update. Consequently, Byz-Clip21-SGD2M is compatible with all of the techniques discussed above and can also be deployed in the central DP setting, where σ_ω^2 scales down with the number of clients.

Error Feedback. To mitigate client drift, which arises from the potentially different local data distributions across clients, we augment our method with an EF21-style error feedback mechanism on the client side. Error feedback controls the discrepancy between the true global update and local updates sent to the server. Incorporating this mechanism into Byz-Clip21-SGD2M allows the algorithm to compensate for the bias introduced by heterogeneous client data, thereby ensuring convergence under larger levels of data heterogeneity. This design choice makes the method robust to one of the central difficulties in federated optimization and allows eliminating the need for the bounded gradient assumption imposed in prior work.

Robust Aggregation. At the heart of robust Byz-Clip21-SGD2M lies the aggregation rule RAgg. To mitigate adversarial clients that attempt to disrupt training, we employ a robust aggregation rule satisfying Definition 3.2. Robust aggregation limits the damage that Byzantine clients can cause during training.

Challenges of Designing Differentially Private and Byzantine-Robust Algorithms. Designing algorithms that are both differentially private and Byzantine-robust

Algorithm 1 Byz-Clip21-SGD2M

Require: $x^0 \in \mathbb{R}^d$, momentum parameters $\beta, \hat{\beta} \in (0, 1]$, step-size $\gamma > 0$, $v_i^0 = g_i^0 = m_i^0 \in \mathbb{R}^d$, clipping parameter $\tau > 0$, DP noise variance $\sigma_\omega^2 \geq 0$

- 1: **for** $t = 0, \dots, T - 1$ **do**
- 2: $x^{t+1} = x^t - \gamma g^t$
- 3: **for** $i \in \mathcal{G}$ **do**
- 4: $v_i^{t+1} = (1 - \beta)v_i^t + \beta \nabla f_i(x^{t+1}, \xi_i^{t+1})$
- 5: $\omega_i^{t+1} \sim \mathcal{N}(0, \sigma_\omega^2 \mathbf{I})$
- 6: $c_i^{t+1} = \text{clip}_\tau(v_i^{t+1} - g_i^t) + \omega_i^{t+1}$
- 7: $g_i^{t+1} = g_i^t + \hat{\beta} \text{clip}_\tau(v_i^{t+1} - g_i^t)$
- 8: **end for**
- 9: **for** $i \in \mathcal{B}$ **do**
- 10: $c_i^{t+1} = (*)$ sends arbitrary vector
- 11: **end for**
- 12: $m_i^{t+1} = m_i^t + \hat{\beta} c_i^{t+1}$ $(*)$ **for** $i \in [n]$
- 13: $g^{t+1} = \text{RAgg}(m_1^{t+1}, \dots, m_n^{t+1})$
- 14: **end for**

(*) This step can be moved to the client side without affecting the DP guarantee and convergence analysis, thanks to the post-processing property [Dwork et al., 2014, Proposition 2.1].

is inherently challenging as the two objectives often conflict with each other. Differential privacy requires adding extra noise to the raw updates, which reduces their consistency and makes it easier for adversarial updates to remain undetected. Moreover, using robust aggregation to combine the DP-noisy updates can amplify the detrimental effects of noise, further hindering convergence. Consequently, theoretical analyses must be carried out with particular care and often under restrictive assumptions. Overcoming these limitations requires carefully designing algorithms that balance privacy, robustness, and practical learnability.

5 THEORETICAL ANALYSIS

Our convergence analysis is based on a carefully designed Lyapunov function

$$\begin{aligned} \Phi^t := & f(x^t) - f^* + \frac{8\gamma\beta}{\hat{\beta}^2\eta^2} \frac{1}{G} \sum_{i \in \mathcal{G}} \|v_i^t - \nabla f_i(x^t)\|^2 \\ & + \frac{2\gamma}{\hat{\beta}\eta} \frac{1}{G} \sum_{i \in \mathcal{G}} \|g_i^t - v_i^t\|^2 + \frac{2\gamma}{\beta} \|\bar{v}^t - \nabla f(x^t)\|^2, \end{aligned} \quad (9)$$

where $\bar{v}^t := \frac{1}{G} \sum_{i \in \mathcal{G}} v_i^t$, and $\eta \sim \tau$, whose value will be specified in the main convergence theorem.

The decrease of this Lyapunov function guarantees convergence of the algorithm, as it implies that $f(x^t) - f^*$ decreases during the training. Moreover, as the iterations proceed, the auxiliary variables v_i^t and g_i^t become increasingly accurate approximations of $\nabla f_i(x^t)$ and v_i^t , respectively. In other words, the momentum buffers provide a smoothed estimate of the true gradients, demonstrating the benefit of

momentum in reducing variance and effectively mitigating the noise introduced during training. The coefficients in the Lyapunov function are carefully chosen to balance its different components, ensuring that each term contributes at the same order and none dominates the others. The detailed proofs are deferred to Appendices D and E.

Theorem 5.1 (Simplified). *Let Assumptions 3.3, 3.4, and 3.5a hold, and $\alpha \in (0, 1)$ be a failure probability. Let $\tilde{B}_{\text{init}} := \max_{i \in \mathcal{G}} \{\|\nabla f_i(x^0)\|\} > 3\tau$, $\eta \sim \tau/\tilde{B}_{\text{init}}$, and $\Delta \geq \Phi^0$ for Φ^0 defined in (9). Then there exists a choice of hyperparameters γ, β , and $\hat{\beta}$ (see a full statement in Theorem D.1) such that the iterates of Byz-Clip21-SGD2M (Alg. 1) run with DP noise variance $\sigma_\omega^2 > 0$ and (c, δ_{byz}) -robust aggregator ($\delta_{\text{byz}} > 0$) satisfy with probability at least $1 - \alpha$ that*

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|^2 \leq \tilde{\mathcal{O}} \left(\Delta_0 (R + R^{2/3}) \right),$$

where

$$R = \frac{\sqrt{d}\sigma_\omega}{\sqrt{GT}\tau} + \frac{\sqrt{c\delta_{\text{byz}}d}\sigma_\omega}{\sqrt{T}\tau} + \sqrt{c\delta_{\text{byz}}}, \quad F^0 := f(x^0) - f^*,$$

$$\begin{aligned} \Delta_0 &:= \sqrt{L\Delta}(\sqrt{L\Delta} + \tilde{B}_{\text{init}} + \sigma) \\ &= \tilde{\mathcal{O}}(LF^0 + \zeta^2 + \sigma(\sqrt{LF^0} + \zeta)), \end{aligned}$$

and $\tilde{\mathcal{O}}$ hides constant and logarithmic factors and higher order terms decreasing in T .³

Remark 5.2. We stress that Assumption 3.5a is invoked only to bound the constant Δ_0 ; the convergence guarantee itself still holds without this condition due to the use of clipping. This general result highlights that bounded heterogeneity is primarily used to control the norms of the updates. In our setting, the updates c_i^{t+1} are used to update the momenta m_i^{t+1} , which are then aggregated through RAgg. The stability of m_i^{t+1} is maintained by the clipping operator, since without such control the norm could otherwise grow unbounded. Moreover, due to our design choice $\hat{\beta} \sim 1/T$, the growth rate of the norm is limited to $\mathcal{O}(t\tau\hat{\beta}) = \mathcal{O}(1)$, ensuring that the momenta remain well-behaved throughout the training process.

Next, we consider the implication of this theorem when σ_ω is set in a special way such that each step of Byz-Clip21-SGD2M satisfies local- (ε, δ) -DP.

Corollary 5.3 (Simplified). *Under the setup of Theorem 5.1, let $\sigma_\omega = \Theta \left(\frac{\tau}{\varepsilon} \sqrt{T \log(\frac{T}{\delta}) \log(\frac{1}{\delta})} \right)$ for some $\varepsilon, \delta \in (0, 1)$. Then there exists a choice of hyperparameters γ, β , and*

³The failure probability α appears in the logarithmic terms similar to prior work [Fang et al., 2019, Bassily et al., 2021, Sadiev et al., 2023].

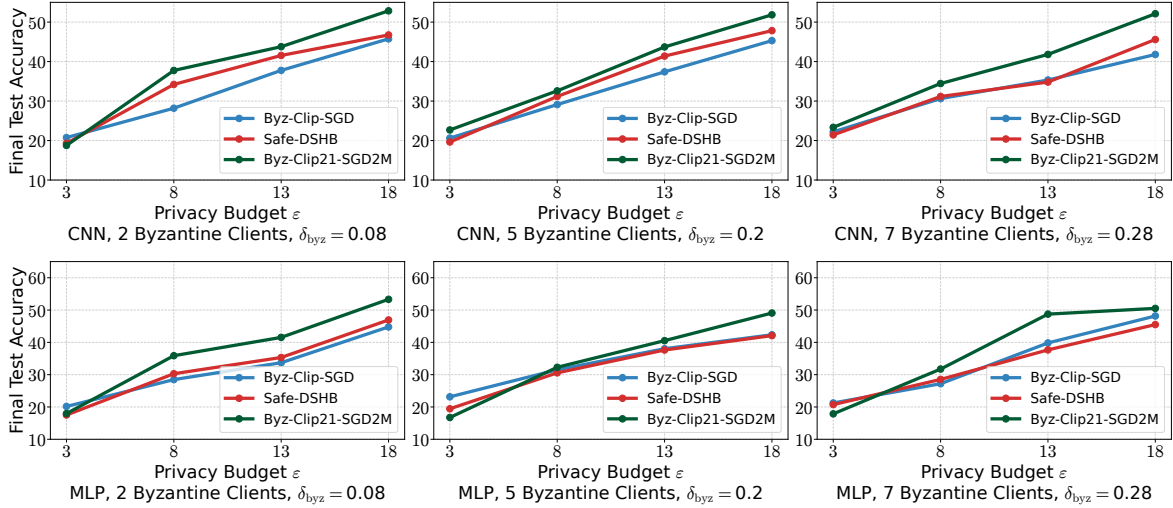


Figure 1: Performance of Byz-Clip21-SGD2M, Byz-Clip-SGD (Algorithm 3), and Safe-DSHB (Algorithm 4) when training CNN (top line) and MLP (bottom line) models on the MNIST dataset for different numbers of Byzantine clients and privacy budgets, when Byzantine clients use IPM attack.

$\hat{\beta}$ (see a full statement in Corollary D.3) such that all T iterations of Byz-Clip21-SGD2M satisfy local (ϵ, δ) -DP with robustness-privacy-utility trade-off bounded as

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|^2 \leq \tilde{\mathcal{O}} \left(\Delta_0(R + R^{2/3}) \right),$$

with probability at least $1 - \alpha$, where

$$R = \frac{\sqrt{d}}{\sqrt{G}\epsilon} + \sqrt{c\delta_{\text{byz}}} \frac{\sqrt{d}}{\epsilon} + \sqrt{c\delta_{\text{byz}}}.$$

In Corollary 5.3, the **first term** corresponds to the error arising from privacy, while the **last term** reflects the error introduced by the presence of Byzantine clients. The **middle term** represents the additional penalty incurred when enforcing both requirements simultaneously. Our utility bound takes a similar form as the lower bound in the strongly convex case [Allouah et al., 2023b], which demonstrates that our analysis is essentially tight. In particular, Theorem 5.1 shows that the impacts of privacy and Byzantine adversaries multiply. The privacy analysis of Byz-Clip21-SGD2M then follows from the well-known result in (Dwork et al. [2014, Theorem 3.22] in combination with the advanced composition theorem ([Dwork et al., 2014, Theorem 3.20]), we obtain that all T iterations of Byz-Clip21-SGD2M are local (ϵ, δ) -differentially private.

Comparison to prior work. We analyze Byz-Clip21-SGD2M under the standard L -smoothness and σ -sub-Gaussian assumptions, thereby improving over earlier work that required the far more restrictive bounded-gradient condition [Zhu and Ling, 2022, Allouah et al., 2023b]. To our knowledge, the only high-probability

study of optimization algorithms in the simultaneous presence of DP and Byzantine adversaries is due to Xiang et al. [2023]; however, their analysis is limited to a single attack model and relies on strong assumptions, thereby limiting its applicability. Other works rely on expectation-based analysis, yet still impose restrictive assumptions; see Table 1 for a concrete comparison. By contrast, our convergence guarantee for Byz-Clip21-SGD2M depends only logarithmically on the failure probability α , which is consistent with standard high-probability analyses, and holds for any aggregation rule satisfying Definition 3.2.

Main Challenges in the Analysis. Server-side aggregation induces a bias between the server aggregate g^t and the honest-average update $\frac{1}{G} \sum_{i \in \mathcal{G}} m_i^t$. This bias scales multiplicatively with the DP noise variance σ_ω and $c\delta_{\text{byz}}$; controlling it without assuming that the gradients are bounded requires a careful analysis with well-calibrated hyperparameters. To address this, we treat clipping as a contractive compressor, with the complication that its contraction factor is input-dependent, which motivates a high-probability analysis. Using a refined inductive argument, we show that the inputs to the clipping operator remain bounded with high probability. Finally, obtaining only poly-logarithmic dependence on the failure probability α hinges on a careful decomposition of the noise terms, each controlled via concentration inequalities.

5.1 CONVERGENCE IN SPECIAL CASES

We now examine several special cases of our bound, which showcase both the tightness of our analysis and the broad applicability of Byz-Clip21-SGD2M. Most follow directly

from Theorem 5.1.

Absence of DP and Byzantine Adversaries. Setting $\sigma_\omega = 0$ and $\delta_{\text{byz}} = 0$ significantly simplifies the choice of hyperparameters. In particular, we can set $\gamma \sim \beta/L$ and $\hat{\beta} = 1$. The following corollary provides a formal convergence in this setting, obtained directly from Theorem 5.1.

Corollary 5.4 (Simplified). *Under assumptions of Theorem 5.1, let $\sigma_\omega = 0$ and $\delta_{\text{byz}} = 0$. Then, there exists a set of hyperparameters $\gamma \sim \beta/L$ and $\hat{\beta} = 1$ such that the iterates of Byz-Clip21-SGD2M satisfy with probability at least $1 - \alpha$ that $\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|^2$ is bounded by*

$$\tilde{\mathcal{O}} \left(\frac{L\Delta(1 + \tilde{B}/\tau)}{T} + \frac{\sigma(\sqrt{L\Delta} + \tilde{B} + \sigma)}{\sqrt{GT}} \right)$$

This result yields the standard $\tilde{\mathcal{O}}(1/\sqrt{GT})$ convergence rate, which matches known lower bounds in this regime [Arjevani et al., 2023]. Importantly, it is established under only the standard assumptions of L -smoothness and σ -sub-Gaussian gradient noise, thereby improving over prior results [Liu et al., 2022, Noble et al., 2022, Allouah et al., 2024b].

DP in the absence of Byzantine Clients. As a next special case, we consider the setting where data is protected solely through DP noise injection, with no Byzantine clients present (i.e., $\delta_{\text{byz}} = 0$). In this regime, the privacy-utility trade-off can be obtained by simply setting $\delta_{\text{byz}} = 0$ in Corollary 5.3.

Corollary 5.5 (Simplified). *Under the setup of Theorem 5.1, let $\sigma_\omega = \Theta\left(\frac{\tau}{\varepsilon} \sqrt{T \log(\frac{T}{\delta}) \log(\frac{1}{\delta})}\right)$ for some $\varepsilon, \delta \in (0, 1)$ and $\delta_{\text{byz}} = 0$. Then there exists a choice of hyperparameters γ, β , and $\hat{\beta}$ such that the iterates of Byz-Clip21-SGD2M satisfy with probability at least $1 - \alpha$ that $\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|^2$ is bounded by*

$$\sqrt{L\Delta}(\sqrt{L\Delta} + \tilde{B} + \sigma) \cdot \tilde{\mathcal{O}} \left(\frac{\sqrt{d}}{\sqrt{G\varepsilon}} + \left(\frac{\sqrt{d}}{\sqrt{G\varepsilon}} \right)^{2/3} \right).$$

In this setting, our general theorem subsumes the results of Islamov et al. [2025b]. Moreover, in modern applications where the model size d is much larger than the number of regular clients G [Charles et al., 2024, Chua et al., 2024], the dominant privacy-utility term is $\tilde{\mathcal{O}}(\sqrt{d}/\sqrt{G\varepsilon})$, matching known lower bounds in this regime [Duchi et al., 2018]. Moreover, unlike prior work, this result does not rely on unrealistic problem assumptions.

Byzantine robustness in the absence of DP noise. In this regime, clipping, originally required for privacy guarantees, is no longer needed. Accordingly, we set $\tau = +\infty$ and $\hat{\beta} = 1$ in Byz-Clip21-SGD2M, which reduces to the

algorithm analyzed by Karimireddy et al. [2020]; see Algorithm 2. This simplification also removes the need to store the server-side momentum buffers $\{m_i^t\}_{i=1}^n$. Since applying Theorem 5.1 directly in this case would yield weaker guarantees, we instead develop a dedicated high-probability analysis, leading to much simpler step-size conditions. Furthermore, the Lyapunov function itself becomes considerably simpler, as the buffers $\{g_i^t\}_{i \in \mathcal{G}}$ are no longer required (see Equation (95)).

Theorem 5.6 (Simplified). *Let Assumption 3.3, Assumption 3.4 and Assumption 3.5 hold. Then there exists a choice of hyperparameters γ, β , and $\hat{\beta} = 1$ such that the iterates of Byz-Clip21-SGD2M with $\sigma_\omega = 0$ satisfy with probability at least $1 - \alpha$ that $\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|^2$ is bounded by*

$$\tilde{\mathcal{O}} \left(\frac{LF^0}{T} + \frac{\sigma(\sqrt{LF^0} + \sigma/\sqrt{G} + \sqrt{c\delta_{\text{byz}}(\zeta + \sigma)}}{\sqrt{GT}} + c\delta_{\text{byz}}\zeta^2 \right).$$

We establish that Byz-Clip21-SGD2M converges at rate $\tilde{\mathcal{O}}(1/\sqrt{GT})$ to a neighborhood of size $\tilde{\mathcal{O}}(c\delta_{\text{byz}}\zeta^2)$, matching the in-expectation lower bound of Karimireddy et al. [2020].

6 EXPERIMENTS

To demonstrate the efficacy of the proposed algorithm, we test the performance of Byz-Clip21-SGD2M against Byz-Clip-SGD (Algorithm 3) and Safe-DSHB (Algorithm 4) [Allouah et al., 2023b] when training CNN and MLP models on the MNIST dataset [LeCun et al., 2010]. We fix the total number of clients to 25, distribute the dataset among them equally, and vary the rate of Byzantine clients (δ_{byz}) that perform IPM (Figure 1) [Xie et al., 2020] or Label Flipping (Figure 2) [Paudice et al., 2018] attacks. Byzantine clients performing the IPM attack compute the average of the update vectors of regular clients and multiply the average by -10 , as described by Xie et al. [2019]. Label flipping is implemented by mapping label y to $9 - y$ (assuming 10 classes numbered from 0).

We evaluate performance by varying the number of Byzantine clients in $\{2, 5, 7\}$ and the privacy budget $\varepsilon \in \{3, 8, 13, 18, 23\}$. For each algorithm, we tune the learning rate parameter over the grid $\{10, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ and the clipping threshold $\tau \in \{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. For Byz-Clip21-SGD2M and Safe-DSHB, we fix the local momentum parameter to $\beta = 0.1$, while we use $\hat{\beta} = 0.01$ for Byz-Clip21-SGD2M. In all setups, we use a batch size of 32. Under this setting, every algorithm allocates an equal amount of privacy budget for tuning, since the set of tested hyperparameters remains the same across all algorithms.

Consistent with our theory, where we assume access to σ -sub-Gaussian stochastic gradients, we disable privacy amplification by subsampling and inject DP noise with standard

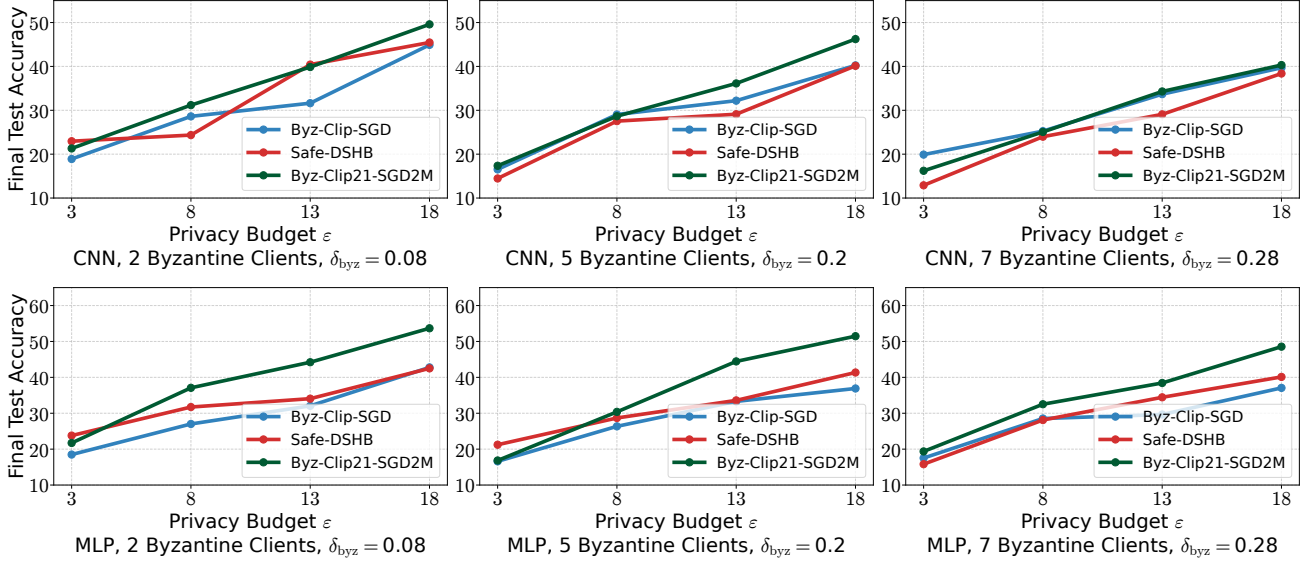


Figure 2: Performance of Byz-Clip21-SGD2M, Byz-Clip-SGD (Algorithm 3), and Safe-DSHB (Algorithm 4) when training CNN (top line) and MLP (bottom line) models on the MNIST dataset for different numbers of Byzantine clients and privacy budgets, when Byzantine clients use label flipping attack.

deviation $\sigma_\omega = \frac{\tau}{\epsilon} \sqrt{T \log(1/\delta)}$, where T is the total number of iterations. Therefore, in this set of experiments, the tested privacy budgets are larger than those in the subsampling experiments of Appendix F.2. In all main experiments, the server uses a coupling of Nearest Neighbor Mixing (NNM) [Allouah et al., 2023a] and Coordinate Median (CM) [Chen et al., 2017] as the aggregation rule, while the Gaussian mechanism utilizes $\delta = \frac{1}{|\mathcal{D}|}$, where $|\mathcal{D}|$ is the dataset size.

In most configurations, Byz-Clip21-SGD2M is competitive, matching or surpassing the baselines in test accuracy. These findings align with our theory and provide empirical evidence that Byz-Clip21-SGD2M is effective for training under DP noise injection and Byzantine attacks. Additional experimental results and training details are summarized in Appendix F.

7 CONCLUSION AND LIMITATIONS

In this work, we present Byz-Clip21-SGD2M, a new algorithm that admits provable convergence under standard L -smoothness and σ -sub-Gaussian gradient noise in the simultaneous presence of DP and Byzantine adversaries. Several directions merit further study: (i) establishing lower bounds without the bounded-gradient assumption of Allouah et al. [2023b]; (ii) tightening our convergence guarantees to match known lower bounds exactly when DP and Byzantine adversaries are considered separately; (iii) extending the analysis of Byz-Clip21-SGD2M to heavy-tailed gradient noise and relaxed smoothness assumptions [Zhang et al., 2020a, Alimisis et al., 2025, Bolatov et al., 2026]; (iv) strengthening the privacy guarantees via amplification by

data and client subsampling.

8 ACKNOWLEDGMENT

The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST): i) KAUST Baseline Research Scheme, ii) CRG Grant ORFS-CRG12-2024-6460, and iii) Center of Excellence for Generative AI, under award number 5940.

References

- Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). Official Journal of the European Union, L119, 1–88, 2016. URL <https://eur-lex.europa.eu/eli/reg/2016/679/oj>. (Cited on page 1)
- Directive (eu) 2022/2555 of the european parliament and of the council of 14 december 2022 on measures for a high common level of cybersecurity across the union (nis2 directive). Official Journal of the European Union, L 333, 27 December 2022, pp. 80–152, 2022. URL <https://eur-lex.europa.eu/eli/dir/2022/2555/oj>. (Cited on page 1)
- Regulation (eu) 2024/2847 of the european parliament and of the council of 13 march 2024 on horizontal

- cybersecurity requirements for products with digital elements (cyber resilience act). Official Journal of the European Union, L 2847, 8 October 2024, 2024. URL <https://eur-lex.europa.eu/eli/reg/2024/2847/oj>. (Cited on page 1)
- Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act). Official Journal of the European Union, L1689, 2024. URL <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>. (Cited on page 1)
- 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*, 2016. (Cited on pages 3 and 63)
- Ahmad Ajalloeian and Sebastian U Stich. On the convergence of sgd with biased gradients. *arXiv preprint arXiv:2008.00051*, 2020. (Cited on page 2)
- Foivos Alimisis, Rustem Islamov, and Aurelien Lucchi. Why do we need warm-up? a theoretical perspective. *arXiv preprint arXiv:2510.03164*, 2025. (Cited on page 9)
- Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in neural information processing systems*, 30, 2017. (Cited on page 2)
- Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine stochastic gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. (Cited on page 2)
- Zeyuan Allen-Zhu, Faeze Ebrahimi, Jerry Li, and Dan Alistarh. Byzantine-resilient non-convex stochastic gradient descent. *arXiv preprint arXiv:2012.14368*, 2020. (Cited on page 2)
- 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*, 2023a. (Cited on pages 2, 3, 4, and 9)
- Youssef Allouah, Rachid Guerraoui, Nirupam Gupta, Rafaël Pinot, and John Stephan. On the privacy-robustness-utility trilemma in distributed learning. In *International Conference on Machine Learning*, 2023b. (Cited on pages 2, 3, 4, 5, 7, 8, 9, and 63)
- Youssef Allouah, Rachid Guerraoui, Nirupam Gupta, Ahmed Jellouli, Geovani Rizk, and John Stephan. Adaptive gradient clipping for robust federated learning. *arXiv preprint arXiv:2405.14432*, 2024a. (Cited on page 2)
- Youssef Allouah, Anastasia Koloskova, Aymane El Firdoussi, Martin Jaggi, and Rachid Guerraoui. The privacy power of correlated noise in decentralized learning. *arXiv preprint arXiv:2405.01031*, 2024b. (Cited on pages 3 and 8)
- Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 2023. (Cited on page 8)
- Raef Bassily, Cristóbal Guzmán, and Michael Menart. Differentially private stochastic optimization: New results in convex and non-convex settings. *Advances in Neural Information Processing Systems*, 2021. (Cited on page 6)
- Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International conference on machine learning*, pages 560–569. PMLR, 2018. (Cited on page 2)
- Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *Journal of Machine Learning Research*, 24(276):1–50, 2023. (Cited on page 2)
- Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 2017. (Cited on pages 1 and 2)
- Arman Bolatov, Samuel Horváth, Martin Takáč, and Eduard Gorbunov. Byzantine-robust optimization under (L_0, L_1) -smoothness. In *The Third Conference on Parsimony and Learning (Proceedings Track)*, 2026. URL <https://openreview.net/forum?id=IhErR5K3EL>. (Cited on page 9)
- 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*, 2017. (Cited on page 3)
- Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 2020. (Cited on page 4)
- Zachary Charles, Arun Ganesh, Ryan McKenna, H Brendan McMahan, Nicole Mitchell, Krishna Pillutla, and Keith Rush. Fine-tuning large language models with user-level differential privacy. *arXiv preprint arXiv:2407.07737*, 2024. (Cited on page 8)
- Kamalika Chaudhuri, Chuan Guo, and Mike Rabbat. Privacy-aware compression for federated data analysis.

- In *Uncertainty in Artificial Intelligence*, 2022. (Cited on page 3)
- Lingjiao Chen, Hongyi Wang, Zachary Charles, and Dimitris Papailiopoulos. Draco: Byzantine-resilient distributed training via redundant gradients. In *International Conference on Machine Learning*, 2018. (Cited on page 2)
- Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2017. (Cited on pages 2 and 9)
- Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Daogao Liu, Pasin Manurangsi, Amer Sinha, and Chiyuan Zhang. Mind the privacy unit! user-level differential privacy for language model fine-tuning. *arXiv preprint arXiv:2406.14322*, 2024. (Cited on page 8)
- Enea Monzio Compagnoni, Alessandro Stanghellini, Rustem Islamov, Aurelien Lucchi, and Anastasiia Koloskova. Adaptive methods are preferable in high privacy settings: An sde perspective. *arXiv preprint arXiv:2603.03226*, 2026. (Cited on page 3)
- Georgios Damaskinos, El Mahdi El Mhamdi, Rachid Guerraoui, Arsany Guirguis, and Sébastien Rouault. Aggregator: Byzantine machine learning via robust gradient aggregation. In *Proceedings of Machine Learning and Systems (MLSys)*, 2019. (Cited on page 2)
- Yury Demidovich, Grigory Malinovsky, Igor Sokolov, and Peter Richtárik. A guide through the zoo of biased sgd. *Advances in Neural Information Processing Systems*, 36: 23158–23171, 2023. (Cited on page 2)
- John C Duchi, Michael I Jordan, and Martin J Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 2018. (Cited on page 8)
- Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 2014. (Cited on pages 4, 6, 7, and 50)
- Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2019. (Cited on page 3)
- Cong Fang, Zhouchen Lin, and Tong Zhang. Sharp analysis for nonconvex sgd escaping from saddle points. In *Conference on Learning Theory*, 2019. (Cited on page 6)
- Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and John Stephan. Byzantine machine learning made easy by resilient averaging of momentums. In *International Conference on Machine Learning*, 2022. (Cited on page 4)
- Ilyas Fatkhullin, Alexander Tyurin, and Peter Richtárik. Momentum provably improves error feedback! *Advances in Neural Information Processing Systems*, 2024. (Cited on pages 2 and 5)
- Ilyas Fatkhullin, Igor Sokolov, Eduard Gorbunov, Zhize Li, and Peter Richtárik. Ef21 with bells & whistles: Six algorithmic extensions of modern error feedback. *Journal of Machine Learning Research*, 26(189):1–50, 2025. (Cited on page 2)
- Xinwen Gao, Shaojing Fu, Lin Liu, and Yuchuan Luo. Bvdfed: Byzantine-resilient and verifiable aggregation for differentially private federated learning. *Frontiers of Computer Science*, 2024. (Cited on page 4)
- Yuan Gao, Rustem Islamov, and Sebastian U Stich. Econtrol: Fast distributed optimization with compression and error control. *arXiv preprint arXiv:2311.05645*, 2023. (Cited on page 2)
- Yuan Gao, Anton Rodomanov, Jeremy Rack, and Sebastian Stich. Composite optimization with error feedback: the dual averaging approach. *arXiv preprint arXiv:2510.03507*, 2025a. (Cited on page 2)
- Yuan Gao, Anton Rodomanov, Jeremy Rack, and Sebastian U Stich. Accelerated distributed optimization with compression and error feedback. *arXiv preprint arXiv:2503.08427*, 2025b. (Cited on page 2)
- Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients — how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems*, 2020. (Cited on page 1)
- Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 2012. (Cited on page 5)
- Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 2013. (Cited on page 4)
- Eduard Gorbunov, Darina Dvinskikh, and Alexander Gasnikov. Optimal decentralized distributed algorithms for stochastic convex optimization. *arXiv preprint arXiv:1911.07363*, 2019. (Cited on page 17)

- Eduard Gorbunov, Dmitry Kovalev, Dmitry Makarenko, and Peter Richtárik. Linearly converging error compensated sgd. *Advances in Neural Information Processing Systems*, 33:20889–20900, 2020. (Cited on page 2)
- Eduard Gorbunov, Alexander Borzunov, Michael Diskin, and Max Ryabinin. Secure distributed training at scale. In *International Conference on Machine Learning*, 2022a. (Cited on page 2)
- 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. *arXiv preprint arXiv:2206.00529*, 2022b. (Cited on pages 2, 3, 4, 5, and 62)
- Xiaolan Gu, Ming Li, and Li Xiong. {DP-BREM}:{Differentially-Private} and {Byzantine-Robust} federated learning with client momentum. In *34th USENIX Security Symposium (USENIX Security 25)*, 2025. (Cited on pages 3 and 4)
- Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, Sebastien Rouault, and John Stephan. Combining differential privacy and byzantine resilience in distributed sgd. *arXiv preprint arXiv:2110.03991*, 2021a. (Cited on page 4)
- Rachid Guerraoui, Nirupam Gupta, Rafaël Pinot, Sébastien Rouault, and John Stephan. Differential privacy and byzantine resilience in sgd: Do they add up? *arXiv preprint arXiv:2102.08166*, 2021b. (Cited on pages 2 and 3)
- Lie He, Sai Praneeth Karimireddy, and Martin Jaggi. Byzantine-robust decentralized learning via clippededgossip. *arXiv preprint arXiv:2202.01545*, 2022. (Cited on page 2)
- Mahmoud Hegazy, Rémi Leluc, Cheuk Ting Li, and Aymeric Dieuleveut. Compression with exact error distribution for federated learning. *arXiv preprint arXiv:2310.20682*, 2023. (Cited on page 3)
- Samuel Horváth and Peter Richtárik. A better alternative to error feedback for communication-efficient distributed learning. *arXiv preprint arXiv:2006.11077*, 2020. (Cited on page 62)
- Xinmeng Huang, Ping Li, and Xiaoyun Li. Stochastic controlled averaging for federated learning with communication compression. *arXiv preprint arXiv:2308.08165*, 2023. (Cited on page 2)
- Rustem Islamov, Xun Qian, Slavomír Hanzely, Mher Safaryan, and Peter Richtárik. Distributed newton-type methods with communication compression and bernoulli aggregation. *Transactions on Machine Learning Research*, 2023. (Cited on page 2)
- Rustem Islamov, Yuan Gao, and Sebastian U Stich. Towards faster decentralized stochastic optimization with communication compression. *arXiv preprint arXiv:2405.20114*, 2024a. (Cited on page 2)
- Rustem Islamov, Mher Safaryan, and Dan Alistarh. As-Grad: A sharp unified analysis of asynchronous-SGD algorithms. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pages 649–657, 2024b. (Cited on page 5)
- Rustem Islamov, Yarden As, and Ilyas Fatkhullin. Safe-ef: Error feedback for nonsmooth constrained optimization. *arXiv preprint arXiv:2505.06053*, 2025a. (Cited on page 2)
- Rustem Islamov, Samuel Horváth, Aurelien Lucchi, Peter Richtárik, and Eduard Gorbunov. Double momentum and error feedback for clipping with fast rates and differential privacy. *arXiv preprint arXiv:2502.11682*, 2025b. (Cited on pages 1, 3, 5, 8, 18, 19, 28, and 53)
- Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. *arXiv preprint arXiv:2006.09365*, 2020. (Cited on pages 2, 3, 4, 8, and 52)
- Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Learning from history for byzantine robust optimization. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. (Cited on pages 2 and 4)
- Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 2011. (Cited on page 3)
- Saleh Vatan Khah, Savelii Chezhegov, Shahrokh Farahmand, Samuel Horváth, and Eduard Gorbunov. Differentially private clipped-sgd: High-probability convergence with arbitrary clipping level. *arXiv preprint arXiv:2507.23512*, 2025. (Cited on page 3)
- Sarit Khirirat, Eduard Gorbunov, Samuel Horváth, Rustem Islamov, Fakhri Karray, and Peter Richtárik. Clip21: Error feedback for gradient clipping. *arXiv preprint arXiv:2305.18929*, 2023. (Cited on page 17)
- Anastasia Koloskova, Tao Lin, Sebastian U Stich, and Martin Jaggi. Decentralized deep learning with arbitrary communication compression. *arXiv preprint arXiv:1907.09356*, 2019. (Cited on page 2)
- Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, 2023. (Cited on page 3)
- Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon.

- Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016. (Cited on page 1)
- Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. (Cited on page 64)
- Muhang Lan, Song Xiao, and Wenyi Zhang. One-bit model aggregation for differentially private and byzantine-robust personalized federated learning. *arXiv preprint arXiv:2507.03973*, 2025. (Cited on page 4)
- Yann LeCun, Corinna Cortes, and C. J. C. Burges. Mnist handwritten digit database. *ATT Labs [Online]*. Available: <http://yann.lecun.com/exdb/mnist>, 2010. (Cited on page 8)
- Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 2020. (Cited on pages 1 and 3)
- Zhize Li, Haoyu Zhao, Boyue Li, and Yuejie Chi. Soteriafl: A unified framework for private federated learning with communication compression. *Advances in Neural Information Processing Systems*, 2022. (Cited on page 1)
- Mingrui Liu, Zhenxun Zhuang, Yunwen Lei, and Chunyang Liao. A communication-efficient distributed gradient clipping algorithm for training deep neural networks. *Advances in Neural Information Processing Systems*, 2022. (Cited on pages 3 and 8)
- Zijian Liu, Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Nguyen. High probability convergence of stochastic gradient methods. In *International Conference on Machine Learning*, 2023. (Cited on page 5)
- Andrew Lowy, Ali Ghafelebashi, and Meisam Razaviyayn. Private non-convex federated learning without a trusted server. In *International Conference on Artificial Intelligence and Statistics*, 2023. (Cited on pages 1 and 3)
- Lingjuan Lyu, Han Yu, Xingjun Ma, Chen Chen, Lichao Sun, Jun Zhao, Qiang Yang, and Philip S Yu. Privacy and robustness in federated learning: Attacks and defenses. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. (Cited on page 1)
- Maksim Makarenko, Elnur Gasanov, Rustem Islamov, Abdurakhmon Sadiev, and Peter Richtárik. Adaptive compression for communication-efficient distributed training. *arXiv preprint arXiv:2211.00188*, 2022. (Cited on page 2)
- Grigory Malinovsky, Peter Richtárik, Samuel Horváth, and Eduard Gorbunov. Byzantine robustness and partial participation can be achieved at once: Just clip gradient differences. *Advances in Neural Information Processing Systems*, 2024. (Cited on pages 2 and 4)
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 2017a. (Cited on page 1)
- H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017b. (Cited on page 3)
- El Mahdi El Mhamdi, Rachid Guerraoui, and Sébastien Rouault. The hidden vulnerability of distributed learning in byzantium. *arXiv preprint arXiv:1802.07927*, 2018. (Cited on page 2)
- Konstantin Mishchenko, Francis Bach, Mathieu Even, and Blake Woodworth. Asynchronous SGD beats minibatch SGD under arbitrary delays. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=4XP0ZuQKXmV>. (Cited on page 5)
- Tomoya Murata and Taiji Suzuki. DIFF2: Differential private optimization via gradient differences for nonconvex distributed learning. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. (Cited on page 3)
- Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 2009. (Cited on page 5)
- Maxence Noble, Aurélien Bellet, and Aymeric Dieuleveut. Differentially private federated learning on heterogeneous data. In *International conference on artificial intelligence and statistics*, 2022. (Cited on pages 3 and 8)
- Shreyas Pai, Gopal Pandurangan, Sriram V Pemmaraju, and Peter Robinson. Can we break symmetry with $o(m)$ communication? In *Proceedings of the 2021 ACM Symposium on Principles of Distributed Computing*, 2021. (Cited on page 4)
- Andrea Paudice, Luis Muñoz-González, and Emil C Lupu. Label sanitization against label flipping poisoning attacks. In *Joint European conference on machine learning and knowledge discovery in databases*, 2018. (Cited on page 8)
- Krishna Pillutla, Sham M Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 2022. (Cited on pages 2 and 4)
- Xun Qian, Rustem Islamov, Mher Safaryan, and Peter Richtárik. Basis matters: better communication-efficient second order methods for federated learning. *arXiv preprint arXiv:2111.01847*, 2021. (Cited on page 2)

- Shashank Rajput, Hongyi Wang, Zachary Charles, and Dimitris Papailiopoulos. Detox: A redundancy-based framework for faster and more robust gradient aggregation. *Advances in neural information processing systems*, 2019. (Cited on page 2)
- Ahmad Rammal, Kaja Grutkowska, Nikita Fedin, Eduard Gorbunov, and Peter Richtárik. Communication compression for byzantine robust learning: New efficient algorithms and improved rates. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024. (Cited on page 2)
- Jayanth Regatti, Hao Chen, and Abhishek Gupta. Bygars: Byzantine sgd with arbitrary number of attackers. *arXiv preprint arXiv:2006.13421*, 2020. (Cited on page 2)
- Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. Ef21: A new, simpler, theoretically better, and practically faster error feedback. *Advances in Neural Information Processing Systems*, 34:4384–4396, 2021. (Cited on page 2)
- Nuria Rodriguez-Barroso, Eugenio Martinez-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. (Cited on page 2)
- 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. In *International Conference on Machine Learning*, 2023. (Cited on page 6)
- Mher Safaryan, Rustem Islamov, Xun Qian, and Peter Richtárik. Fednl: Making newton-type methods applicable to federated learning. *arXiv preprint arXiv:2106.02969*, 2021. (Cited on page 2)
- 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 *Interspeech*, volume 2014, pages 1058–1062. Singapore, 2014. (Cited on page 2)
- Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, 2017. (Cited on page 1)
- Egor Shulgin, Sarit Khirirat, and Peter Richtárik. Smoothed normalization for efficient distributed private optimization. *arXiv preprint arXiv:2502.13482*, 2025a. (Cited on page 3)
- Egor Shulgin, Grigory Malinovsky, Sarit Khirirat, and Peter Richtárik. First provable guarantees for practical private fl: Beyond restrictive assumptions. *Tiny Titans: The Next Wave of On-Device Learning for Foundational Models (ICML workshop)*, 2025b. (Cited on pages 1 and 3)
- Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019. (Cited on page 2)
- Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. *Advances in neural information processing systems*, 31, 2018. (Cited on page 2)
- Elham Tabassi et al. Artificial intelligence risk management framework (ai rmf 1.0). Technical report, National Institute of Standards and Technology, 2023. URL <https://doi.org/10.6028/NIST.AI.100-1>. (Cited on page 1)
- Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction APIs. In *USENIX Security Symposium*, 2016. (Cited on page 1)
- Lingxiao Wang, Bargav Jayaraman, David Evans, and Quanquan Gu. Efficient privacy-preserving stochastic nonconvex optimization. In *Uncertainty in Artificial Intelligence*, 2023. (Cited on pages 1 and 3)
- Lingxiao Wang, Xingyu Zhou, Kumar Kshitij Patel, Lawrence Tang, and Aadirupa Saha. Efficient private federated non-convex optimization with shuffled model. In *Privacy Regulation and Protection in Machine Learning*, 2024. (Cited on page 3)
- Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. *Advances in Neural Information Processing Systems*, 31, 2018. (Cited on page 2)
- 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*, 2020. (Cited on pages 2 and 5)
- Zihang Xiang, Tianhao Wang, Wanyu Lin, and Di Wang. Practical differentially private and byzantine-resilient federated learning. *Proceedings of the ACM on Management of Data*, 2023. (Cited on pages 2, 3, 4, and 7)
- Cong Xie, Sanmi Koyejo, and Indranil Gupta. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance. In *International conference on machine learning*, 2019. (Cited on page 8)
- Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation. In *Uncertainty in artificial intelligence*, 2020. (Cited on page 8)

- Xinyi Xu and Lingjuan Lyu. A reputation mechanism is all you need: Collaborative fairness and adversarial robustness in federated learning. *arXiv preprint arXiv:2011.10464*, 2020. (Cited on page 2)
- Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2019. (Cited on page 1)
- Chung-Yiu Yau and Hoi-To Wai. Docom: Compressed decentralized optimization with near-optimal sample complexity. *arXiv preprint arXiv:2202.00255*, 2022. (Cited on page 2)
- Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International conference on machine learning*, 2018. (Cited on pages 2 and 4)
- Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. *Advances in Neural Information Processing Systems*, 33:15511–15521, 2020a. (Cited on page 9)
- Meifan Zhang, Zhanhong Xie, and Lihua Yin. Private and communication-efficient federated learning based on differentially private sketches. *arXiv preprint arXiv:2410.05733*, 2024. (Cited on page 3)
- Xin Zhang, Minghong Fang, Jia Liu, and Zhengyuan Zhu. Private and communication-efficient edge learning: A sparse differential gaussian-masking distributed SGD approach. In *Proceedings of the Twenty-First International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, 2020b. (Cited on page 3)
- Zikai Zhang and Rui Hu. Byzantine-robust federated learning with variance reduction and differential privacy. In *2023 IEEE Conference on Communications and Network Security (CNS)*, 2023. (Cited on page 2)
- Puning Zhao, Jiafei Wu, Zhe Liu, Chong Wang, Rongfei Fan, and Qingming Li. Differential private stochastic optimization with heavy-tailed data: Towards optimal rates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. (Cited on page 3)
- Heng Zhu and Qing Ling. Bridging differential privacy and byzantine-robustness via model aggregation. *arXiv preprint arXiv:2205.00107*, 2022. (Cited on page 7)
- Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 2019. (Cited on page 1)

CONTENTS

A	NOTATION	16
B	USEFUL LEMMAS	17
C	DESCENT LEMMAS	18
D	PROOF OF THEOREM 5.1	28
E	PROOFS IN SPECIAL CASES	50
E.1	Convergence in the Absence of Both DP and Byzantine Adversaries	50
E.2	Convergence in the Absence of Byzantine Adversaries	51
E.3	Convergence in the Absence of DP Adversaries	52
F	ADDITIONAL EXPERIMENTS AND TRAINING DETAILS	62
F.1	Training Details	63
F.2	Amplification by Sub-sampling	63

A NOTATION

In our analysis, we make use of the following quantities

$$\begin{aligned} \bar{v}^t &:= \frac{1}{G} \sum_{i \in \mathcal{G}} v_i^t, \quad \bar{g}^t := \frac{1}{G} \sum_{i \in \mathcal{G}} g_i^t, \quad \bar{m}^t := \frac{1}{G} \sum_{i \in \mathcal{G}} m_i^t, \quad \Omega_i^t := \sum_{l=1}^t \omega_i^l, \quad \bar{\Omega}^t := \frac{1}{G} \sum_{i \in \mathcal{G}} \Omega_i^t, \quad \theta^t := \frac{1}{G} \sum_{i \in \mathcal{G}} \theta_i^t, \\ \nabla f(x^t, \xi^t) &:= \frac{1}{G} \sum_{i \in \mathcal{G}} \nabla f_i(x^t, \xi_i^t). \end{aligned} \quad (10)$$

Next, we derive the relation between $\bar{m}^t$ and $\bar{g}^t$ of the form

$$\bar{m}^t = \bar{g}^t + \hat{\beta} \Omega^t \quad (11)$$

Indeed, we have $\bar{m}^0 = \bar{g}^0$ by initialization of Algorithm 1. Assume that the relation holds at iteration t , let us show that it also holds at iteration $t + 1$:

$$\begin{aligned} \bar{m}^{t+1} &= \bar{m}^t + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} [\text{clip}_\tau(v_i^{t+1} - g_i^t) + \omega_i^{t+1}] \\ &\stackrel{\text{Ind. Asmp.}}{=} \bar{g}^t + \hat{\beta} \bar{\Omega}^t + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} \text{clip}_\tau(v_i^{t+1} - g_i^t) + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} \omega_i^{t+1} \\ &\stackrel{(10)}{=} \frac{1}{G} \sum_{i \in \mathcal{G}} [g_i^t + \hat{\beta} \text{clip}_\tau(v_i^{t+1} - g_i^t)] + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} [\Omega_i^t + \omega_i^{t+1}] \\ &= \frac{1}{G} \sum_{i \in \mathcal{G}} g_i^{t+1} + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} \Omega_i^{t+1} \\ &= \bar{g}^{t+1} + \hat{\beta} \bar{\Omega}^{t+1}. \end{aligned} \quad (12)$$

Similarly, we obtain

$$m_i^t = g_i^t + \hat{\beta} \Omega_i^t. \quad (13)$$

Table 2: Summary of notation used in the paper, including convergence proofs.

Symbol	Definition	Meaning	Reference
n	—	Number of all clients	Section 3
x	—	Model parameters	Section 3
d	—	Number of model parameters	Section 3
$\mathcal{G}$	—	Set of regular clients	Section 3
G	—	Number of regular clients	Section 3
$\mathcal{B}$	—	Set of Byzantine clients	Section 3
δ_{byz}	—	Fraction of Byzantine clients	Section 3
ε, δ	—	Privacy budget constants	Definition 3.1
τ	—	Clipping threshold	Equation (8)
c	—	Scaling constant in aggregation rule	Definition 3.2
L	—	Smoothness constant	Assumption 3.3
A_ζ, ζ^2	$\frac{1}{G} \sum_{i \in \mathcal{G}} \ \nabla f_i(x) - \nabla f(x)\ ^2 \leq A_\zeta \ \nabla f(x)\ ^2 + \zeta^2$ or $\max_{i \in \mathcal{G}} \ \nabla f_i(x) - \nabla f(x)\ ^2 \leq A_\zeta \ \nabla f(x)\ ^2 + \zeta^2$	Heterogeneity bound	Assumption 3.5 or Assumption 3.5a
f^*	$f^* := \inf_{x \in \mathbb{R}^d} f(x) > -\infty$	Lower bound on the function value	Assumption 3.3
T	—	Number of iterations	—
$\nabla f_i(x^t, \xi_i^t)$	—	Stochastic gradient at iteration t of client $i \in \mathcal{G}$	Assumption 3.4
$\nabla f(x^t, \xi^t)$	$\nabla f(x^t, \xi^t) := \frac{1}{G} \sum_{i \in \mathcal{G}} \nabla f_i(x^t, \xi_i^t)$	Averaged stochastic gradient at iteration t of all regular clients	Equation (10)
$\bar{v}^t$	$\bar{v}^t := \frac{1}{G} \sum_{i \in \mathcal{G}} v_i^t$	Averaged client momentum buffer	Equation (10)
$\bar{g}^t$	$\bar{g}^t := \frac{1}{G} \sum_{i \in \mathcal{G}} g_i^t$	Averaged EF21 learnable shifted vector	Equation (10)
$\bar{m}^t$	$\bar{m}^t := \frac{1}{G} \sum_{i \in \mathcal{G}} m_i^t$	Averaged server momentum buffer	Equation (10)
ω_i^t	—	DP noise added by client $i \in \mathcal{G}$ at iteration t	Algorithm 1
Ω_i^t	$\Omega_i^t := \sum_{l=1}^t \omega_i^l, \quad \bar{\Omega}^t := \frac{1}{G} \sum_{i \in \mathcal{G}} \Omega_i^t$	Accumulated DP noise of client $i \in \mathcal{G}$	Equation (10)
$\bar{\Omega}^t$	$\bar{\Omega}^t := \frac{1}{G} \sum_{i \in \mathcal{G}} \Omega_i^t$	Averaged accumulated DP noise of all regular clients	Equation (10)
θ_i^t	$\theta_i^t := \nabla f_i(x^t, \xi_i^t) - \nabla f_i(x^t)$	Mini-batch noise at iteration t of regular client $i \in \mathcal{G}$	Assumption 3.4
θ^t	$\theta^t := \frac{1}{G} \sum_{i \in \mathcal{G}} \theta_i^t$	Averaged mini-batch noise of all regular clients	Equation (10)
Φ^t	—	Lyapunov function	Equation (9)
$\tilde{\Phi}^t$	—	Lyapunov function in the absence of DP adversaries	Equation (95)
α	—	Failure probability	Theorem 5.1
Δ	$\Delta \geq \Phi^0$	Upper bound on the Lyapunov function at initialization	Lemma C.2
B_{init}	$B_{\text{init}} := \max\{3\tau, \max_i\{\ \nabla f_i(x^0)\ \} + b\}$	Sub-optimality of initialization	Theorem D.1
a	$a := \left(\sqrt{2} + 2\sqrt{3 \log \frac{8(T+1)}{\alpha}}\right) \sqrt{d} \sigma_\omega \sqrt{\frac{T}{G}}$	High-probability bound on Ω^t	Equation (67)
$\hat{a}$	$\hat{a} := \left(\sqrt{2} + 2\sqrt{3 \log \frac{8(T+1)}{\alpha}}\right) \sqrt{d} \sigma_\omega \sqrt{T}$	High-probability bound on Ω_i^t	Equation (67)
b^2	$b^2 := 2\sigma^2 \log \frac{16(T+1)G}{\alpha}$	High-probability bound on $\ \theta_i^t\ $	Equation (67)
$\hat{c}^2$	$\hat{c}^2 := \left(\sqrt{2} + 2\sqrt{3 \log \frac{8(T+1)}{\alpha}}\right)^2 \sigma^2$	High-probability bound on $\ \theta^t\ $	Equation (67)
$\tilde{b}^2$	$\tilde{b}^2 := 2\sigma^2 \log \left(\frac{6(T+1)G}{\alpha}\right)$	High-probability bound on $\ \theta_i^t\ $ (in the absence of DP adversaries)	Equation (96)
$\tilde{c}^2$	$\tilde{c}^2 := \left(\sqrt{2} + 2\sqrt{3 \log \frac{8(T+1)}{\alpha}}\right) \sigma^2$	High-probability bound on $\ \theta^t\ $ (in the absence of DP adversaries)	Equation (96)
$\tilde{z}^2$	$\tilde{z}^2 := \left(\sqrt{2} + 2\sqrt{3 \log \frac{8G(T+1)}{\alpha}}\right) \sigma^2$	High-probability bound on $\ \sum_{k=0}^t (1-\beta)^{t-k} (\theta^k - \theta_i^k)\ $ (in the absence of DP adversaries)	Equation (96)

B USEFUL LEMMAS

First, we start with a key property of the clipping operator.

Lemma B.1 (Lemma 4.1 in [Khairat et al., 2023]). *The clipping operator satisfies for any $x \in \mathbb{R}^d$*

$$\|\text{clip}_\tau(x) - x\| \leq \max\{\|x\| - \tau, 0\}. \quad (14)$$

Next, in our high probability analysis, we make use of the following concentration inequality.

Lemma B.2 (Lemma C.3 in [Gorbunov et al., 2019]). *Let $\{\xi_k\}_{k=1}^N$ be the sequence of random vectors with values in $\mathbb{R}^n$ such that*

$$\mathbb{E}[\xi_k \mid \xi_{k-1}, \dots, \xi_1] = 0 \text{ almost surely, } \forall k \in \{1, \dots, N\},$$

and set $S_N := \sum_{k=1}^N \xi_k$. Assume that the sequence $\{\xi_k\}_{k=1}^N$ are sub-Gaussian, i.e.

$$\mathbb{E}[\exp(\|\xi_k\|^2/\sigma_k^2 \mid \xi_{k-1}, \dots, \xi_1)] \leq \exp(1) \text{ almost surely, } \forall k \in \{1, \dots, N\},$$

where $\sigma_2, \dots, \sigma_N$ are some positive numbers. Then for all $\gamma \geq 0$

$$\mathbb{P} \left(\|S_N\| \geq (\sqrt{2} + 2\gamma) \sqrt{\sum_{k=1}^N \sigma_k^2} \right) \leq \exp(-\gamma^2/3). \quad (15)$$

C DESCENT LEMMAS

Lemma C.1. *Let f be L -smooth, $F^t := f(x^t) - f^*$, $\{x^t\}$ be generated by Algorithm 1 with $\gamma \leq \frac{1}{2L}$. Then we have*

$$\begin{aligned} f(x^{t+1}) &\leq f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \frac{1}{4\gamma} \|x^{t+1} - x^t\|^2 + 2\gamma \|\nabla f(x^t) - \bar{v}^t\|^2 + \frac{2\gamma}{G} \sum_{i \in \mathcal{G}} \|v_i^t - g_i^t\|^2 \\ &\quad + 2\gamma \hat{\beta}^2 \|\Omega^t\|^2 + 16\gamma \hat{\beta}^2 c\delta_{\text{byz}} \tau^2 T^2 + 16\gamma \hat{\beta}^2 \frac{c\delta_{\text{byz}}}{G} \sum_{i \in \mathcal{G}} \|\Omega_i^t\|^2, \end{aligned} \quad (16)$$

Proof. Using the derivations from [Islamov et al., 2025b], Lemma 2 we first get

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

We continue the derivations as follows

$$\begin{aligned} f(x^{t+1}) &\stackrel{(i)}{\leq} f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \frac{\gamma}{4} \|g^t\|^2 + 2\gamma \|\nabla f(x^t) - \bar{v}^t\|^2 + 2\gamma \|\bar{v}^t - \bar{g}^t\|^2 + 2\gamma \|\bar{g}^t - \bar{m}^t\|^2 \\ &\quad + 2\gamma \|\bar{m}^t - g^t\|^2 \\ &\stackrel{(ii)}{\leq} f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \frac{1}{4\gamma} \|x^{t+1} - x^t\|^2 + 2\gamma \|\nabla f(x^t) - \bar{v}^t\|^2 + \frac{2\gamma}{G} \sum_{i \in \mathcal{G}} \|v_i^t - g_i^t\|^2 \\ &\quad + 2\gamma \hat{\beta}^2 \|\Omega^t\|^2 + 2\gamma \|\bar{m}^t - g^t\|^2, \end{aligned} \quad (17)$$

where (i) follows from Jensen's inequality applied to $\|\cdot\|^2$, (ii) is obtained using again Jensen's inequality applied to $\|\cdot\|^2$, the update rule of x^t , and (11). Now we bound the term $\|\bar{m}^t - g^t\|^2$ using the properties of the aggregator. We have

$$\begin{aligned} \|\bar{m}^t - g^t\|^2 &= \|\bar{m}^t - \text{RAgg}(m_1^t, \dots, m_n^t)\|^2 \\ &\stackrel{(iii)}{\leq} \frac{c\delta_{\text{byz}}}{G} \sum_{i \in \mathcal{G}} \|\bar{m}^t - m_i^t\|^2 \\ &\stackrel{(iv)}{\leq} \frac{c\delta_{\text{byz}}}{G^2} \sum_{j, l \in \mathcal{G}} \|m_j^t - m_l^t\|^2 = \frac{c\delta_{\text{byz}}}{G^2} \sum_{j, l \in \mathcal{G}, j \neq l} \|m_j^t - m_l^t\|^2 \\ &\leq \frac{c\delta_{\text{byz}}}{G(G-1)} \sum_{j, l \in \mathcal{G}, j \neq l} \|m_j^t - m_l^t\|^2 \\ &\stackrel{(v)}{=} \frac{c\delta_{\text{byz}}}{G(G-1)} \sum_{j, l \in \mathcal{G}, j \neq l} \|g_j^t + \hat{\beta}\Omega_j^t - g_l^t - \hat{\beta}\Omega_l^t\|^2 \\ &\stackrel{(vi)}{\leq} \frac{8c\delta_{\text{byz}}}{G} \sum_{i \in \mathcal{G}} \left\| \hat{\beta} \sum_{s=0}^t \text{clip}_\tau(v_i^s - g_i^{s-1}) \right\|^2 + \frac{8c\delta_{\text{byz}}\hat{\beta}^2}{G} \sum_{i \in \mathcal{G}} \|\Omega_i^t\|^2 \\ &\stackrel{(vii)}{\leq} 8c\delta_{\text{byz}}\hat{\beta}^2(t+1)^2\tau^2 + \frac{8c\delta_{\text{byz}}\hat{\beta}^2}{G} \sum_{i \in \mathcal{G}} \|\Omega_i^t\|^2, \end{aligned} \quad (18)$$

where (iii) follows from the definition of RAgg; (iv) – from Young's inequality, (v) – from (13), (vi) – from Young's inequality and update rule of g_i^t together with the initialization $g_i^0 = 0$, (vii) – from Young's inequality. Plugging (18) in

(17) and noting that $(t+1) \leq T$, we get

$$\begin{aligned} f(x^{t+1}) &\leq f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \frac{1}{4\gamma} \|x^{t+1} - x^t\|^2 + 2\gamma \|\nabla f(x^t) - \bar{v}^t\|^2 + \frac{2\gamma}{G} \sum_{i \in \mathcal{G}} \|v_i^t - g_i^t\|^2 \\ &\quad + 2\gamma \hat{\beta}^2 \|\Omega^t\|^2 + 16\gamma \hat{\beta}^2 c\delta_{\text{byz}} \tau^2 T^2 + 16\gamma \hat{\beta}^2 \frac{c\delta_{\text{byz}}}{G} \sum_{i \in \mathcal{G}} \|\Omega_i^t\|^2, \end{aligned} \quad (19)$$

that finalizes the proof. $\square$

Lemma C.2. *Let each f_i be L -smooth. Then, for the iterates of Algorithm 1 we have the following inequality with probability 1*

$$\begin{aligned} \|v_i^{t+1} - g_i^t\| &\leq (1 - \hat{\beta}) \|v_i^t - g_i^{t-1}\| + \hat{\beta} \max\{0, \|v_i^t - g_i^{t-1}\| - \tau\} + \beta L\gamma \|\bar{m}^t\| + \beta \|\nabla f_i(x^t) - v_i^t\| \\ &\quad + 2\beta \hat{\beta} \tau T L \gamma \sqrt{2c\delta_{\text{byz}}} + 2\beta \hat{\beta} L \gamma \sqrt{\frac{2c\delta_{\text{byz}}}{G} \sum_{i \in \mathcal{G}} \|\Omega_i^t\|^2} + \beta \|\theta_i^{t+1}\|, \end{aligned}$$

where $\theta_i^t := \nabla f_i(x^t, \xi_i^t) - \nabla f_i(x^t)$.

Proof. First, we follow the steps of the proof of Islamov et al. [2025b, Lemma 12] to derive

$$\begin{aligned} \|v_i^{t+1} - g_i^t\| &\leq (1 - \hat{\beta}) \|v_i^t - g_i^{t-1}\| + \hat{\beta} \max\{0, \|v_i^t - g_i^{t-1}\| - \tau\} + \beta L\gamma \|g^t\| + \beta \|\nabla f_i(x^t) - v_i^t\| \\ &\quad + \beta \|\theta_i^{t+1}\|. \end{aligned} \quad (20)$$

Next, we continue as follows

$$\begin{aligned} \|v_i^{t+1} - g_i^t\| &\leq (1 - \hat{\beta}) \|v_i^t - g_i^{t-1}\| + \hat{\beta} \max\{0, \|v_i^t - g_i^{t-1}\| - \tau\} + \beta L\gamma \|g^t - \bar{m}^t\| + \beta L\gamma \|\bar{m}^t\| \\ &\quad + \beta \|\nabla f_i(x^t) - v_i^t\| + \beta \|\theta_i^{t+1}\|. \end{aligned} \quad (21)$$

Now we perform similar derivations as in Lemma C.1 to bound $\|g^t - \bar{m}^t\|$ (see (18)), which finally gives the bound

$$\begin{aligned} \|v_i^{t+1} - g_i^t\| &\leq (1 - \hat{\beta}) \|v_i^t - g_i^{t-1}\| + \hat{\beta} \max\{0, \|v_i^t - g_i^{t-1}\| - \tau\} + \beta L\gamma \|\bar{m}^t\| + \beta \|\nabla f_i(x^t) - v_i^t\| \\ &\quad + \beta L\gamma \sqrt{8c\delta_{\text{byz}} \hat{\beta}^2 T^2 \tau^2 + 8\hat{\beta}^2 \frac{c\delta_{\text{byz}}}{G} \sum_{i \in \mathcal{G}} \|\Omega_i^t\|^2} + \beta \|\theta_i^{t+1}\| \\ &\stackrel{(i)}{\leq} (1 - \hat{\beta}) \|v_i^t - g_i^{t-1}\| + \hat{\beta} \max\{0, \|v_i^t - g_i^{t-1}\| - \tau\} + \beta L\gamma \|\bar{m}^t\| + \beta \|\nabla f_i(x^t) - v_i^t\| \\ &\quad + 2\beta \hat{\beta} \tau T L \gamma \sqrt{2c\delta_{\text{byz}}} + 2\beta \hat{\beta} L \gamma \sqrt{\frac{2c\delta_{\text{byz}}}{G} \sum_{i \in \mathcal{G}} \|\Omega_i^t\|^2} + \beta \|\theta_i^{t+1}\|, \end{aligned} \quad (22)$$

where (i) follows from the triangle inequality: $\sqrt{s+q} \leq \sqrt{s} + \sqrt{q}$ for any $s, q \geq 0$. $\square$

Lemma C.3. *Let each f_i be L -smooth, and $\Delta \geq \Phi^0$. Assume that the following inequalities hold for the iterates generated by Byz-Clip21-SGD2M*

1. $\gamma \leq \frac{1}{24L}$,
2. $\beta \in [0, 1]$, $\hat{\beta} \in [0, 1/24]$,
3. $\|\bar{m}^{t-1}\| \leq \sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\hat{\beta}a + \hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T$,
4. $\|\bar{g}^{t-1}\| \leq \sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T$;
5. $\|\nabla f_i(x^{t-1}) - v_i^{t-1}\| \leq \sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T$ for all $i \in \mathcal{G}$;
6. $\|v_i^t - g_i^{t-1}\| \leq B_{\text{init}}$ for all $i \in \mathcal{G}$;
7. $\|\Omega^t\| \leq a$;
8. $\|\Omega_i^{t-1}\| \leq \hat{a}$ for all $i \in \mathcal{G}$;

9. $\|\theta_i^t\| \leq b$ for all $i \in \mathcal{G}$;
 10. $\Phi^{t-1} \leq 2\Delta$.

Then we have

$$\|\bar{m}^t\| \leq \sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\hat{\beta}a + \hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T. \quad (23)$$

Proof. We start as follows

$$\begin{aligned} \|\bar{m}^t\| &\stackrel{(i)}{=} \left\| \bar{m}^{t-1} + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} [\text{clip}_\tau(v_i^t - g_i^{t-1}) + \omega_i^t] \right\| \\ &= \left\| \bar{m}^{t-1} + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} [\nabla f_i(x^t) + (v_i^t - \nabla f_i(x^t)) + \text{clip}_\tau(v_i^t - g_i^{t-1}) - (v_i^t - g_i^{t-1})] \right. \\ &\quad \left. - \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} g_i^{t-1} + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} \omega_i^t \right\| \\ &= \left\| \bar{m}^{t-1} + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} [\nabla f_i(x^t) + (v_i^t - \nabla f_i(x^t)) + \text{clip}_\tau(v_i^t - g_i^{t-1}) - (v_i^t - g_i^{t-1})] \right. \\ &\quad \left. - \bar{g}^{t-1} + (1 - \hat{\beta})\bar{g}^{t-1} + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} \omega_i^t \right\|. \end{aligned}$$

where (i) follows from the update rule of $\bar{m}^t$. We continue the derivations as follows

$$\begin{aligned} \|\bar{m}^t\| &\stackrel{(ii)}{\leq} \left\| \bar{m}^{t-1} - \bar{g}^{t-1} + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} \omega_i^t \right\| + \hat{\beta}\|\nabla f(x^t)\| + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} \|\text{clip}_\tau(v_i^t - g_i^{t-1}) - (v_i^t - g_i^{t-1})\| \\ &\quad + (1 - \hat{\beta})\|\bar{g}^{t-1}\| + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} \|v_i^t - \nabla f_i(x^t)\| \\ &\stackrel{(iii)}{\leq} \left\| \bar{g}^{t-1} + \hat{\beta}\Omega^{t-1} - \bar{g}^{t-1} + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} \omega_i^t \right\| + \hat{\beta}\|\nabla f(x^{t-1})\| + \hat{\beta}\|\nabla f(x^{t-1}) - \nabla f(x^t)\| \\ &\quad + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} \|\text{clip}_\tau(v_i^t - g_i^{t-1}) - (v_i^t - g_i^{t-1})\| + (1 - \hat{\beta})\|\bar{g}^{t-1}\| \\ &\quad + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} \|(1 - \beta)v_i^{t-1} + \beta\nabla f_i(x^t, \xi_i^t) - \nabla f_i(x^t)\|, \end{aligned} \quad (24)$$

where (ii) follows from the triangle inequality, (iii) – from (11), triangle inequality, and update rule of v_i^t . Using the definition of Ω^t we continue as follows

$$\begin{aligned} \|\bar{m}^t\| &\stackrel{(iv)}{\leq} \hat{\beta}\|\Omega^t\| + \hat{\beta}\|\nabla f(x^{t-1})\| + \hat{\beta}L\gamma\|g^{t-1}\| + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} \max\{0, \|v_i^t - g_i^{t-1}\| - \tau\} + (1 - \hat{\beta})\|\bar{g}^{t-1}\| \\ &\quad + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} ((1 - \beta)\|v_i^{t-1} - \nabla f_i(x^t)\| + \beta\|\nabla f_i(x^t, \xi_i^t) - \nabla f_i(x^t)\|) \\ &\stackrel{(v)}{\leq} \hat{\beta}\|\Omega^t\| + \hat{\beta}\sqrt{2L(f(x^{t-1}) - f^*)} + (2 - \beta)\hat{\beta}L\gamma\|g^{t-1}\| + \hat{\beta}(B_{\text{init}} - \tau) + (1 - \hat{\beta})\|\bar{g}^{t-1}\| \\ &\quad + \frac{\hat{\beta}\beta}{G} \sum_{i \in \mathcal{G}} \|\theta_i^t\| + \frac{\hat{\beta}}{G} (1 - \beta) \sum_{i \in \mathcal{G}} \|v_i^{t-1} - \nabla f_i(x^{t-1})\| \\ &\stackrel{(vi)}{\leq} \hat{\beta}\|\Omega^t\| + \hat{\beta}\sqrt{2L\Phi^{t-1}} + 2(2 - \beta)\hat{\beta}L\gamma\|\bar{m}^{t-1} - g^{t-1}\| + 2(2 - \beta)\hat{\beta}L\gamma\|\bar{m}^{t-1}\| + \hat{\beta}(B_{\text{init}} - \tau) \\ &\quad + (1 - \hat{\beta})\|\bar{g}^{t-1}\| + \frac{\hat{\beta}\beta}{G} \sum_{i \in \mathcal{G}} \|\theta_i^t\| + \frac{\hat{\beta}}{G} (1 - \beta) \sum_{i \in \mathcal{G}} \|v_i^{t-1} - \nabla f_i(x^{t-1})\|, \end{aligned}$$

where (iv) follows from L -smoothness, update rule of x^t , the clipping property from Lemma B.1, and triangle inequality, (v) – from L -smoothness, triangle inequality, the update rule of x^t , the definition of θ_i^t , and the assumption 5 of the lemma, (vi) – from the definition of Φ^t and triangle inequality. Now we use (18) and the assumptions of the lemma to derive

$$\begin{aligned}
\|\bar{m}^t\| &\stackrel{(vii)}{\leq} \hat{\beta}a + \hat{\beta}\sqrt{4L\Delta} + 4\hat{\beta}L\gamma\sqrt{8c\delta_{\text{byz}}\hat{\beta}^2T^2\tau^2 + \hat{\beta}^2\frac{8c\delta_{\text{byz}}}{G}\sum_{i\in\mathcal{G}}\|\Omega_i^{t-1}\|^2} \\
&\quad + 4\hat{\beta}L\gamma\left(\sqrt{4L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\hat{\beta}a + \hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T\right) + \hat{\beta}(B_{\text{init}} - \tau) \\
&\quad + (1 - \hat{\beta})\left(\sqrt{4L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T\right) + \hat{\beta}\beta b \\
&\quad + \hat{\beta}(1 - \beta)\left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T\right) \\
&\stackrel{(viii)}{\leq} \hat{\beta}a + \hat{\beta}\sqrt{4L\Delta} + 8\hat{\beta}^2T\tau L\gamma\sqrt{2c\delta_{\text{byz}}} + 8\hat{\beta}^2L\gamma\sqrt{2c\delta_{\text{byz}}}\hat{a} + \hat{\beta}(B_{\text{init}} - \tau) + \hat{\beta}\beta b \\
&\quad + 4\hat{\beta}L\gamma\left(\sqrt{4L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\hat{\beta}a + 3\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T\right) \\
&\quad + (1 - \hat{\beta})\left(\sqrt{4L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T\right) \\
&\quad + \hat{\beta}(1 - \beta)\left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T\right), \tag{25}
\end{aligned}$$

where (vii) follows from assumptions 3-5, 7, 9, 10 (viii) – from inequality $\sqrt{s+q} \leq \sqrt{s} + \sqrt{q}$ for any $a, b \geq 0$. Rearranging terms, we derive

$$\begin{aligned}
\|\bar{m}^t\| &\leq \sqrt{L\Delta}[2\hat{\beta} + 32\hat{\beta}L\gamma + 8(1 - \hat{\beta}) + 2\hat{\beta}(1 - \beta)] + (B_{\text{init}} - \tau)[12\hat{\beta}L\gamma + 3(1 - \hat{\beta}) + \hat{\beta} + \frac{3}{2}\hat{\beta}(1 - \beta)] \\
&\quad + b[12\hat{\beta}L\gamma + 3(1 - \hat{\beta}) + \hat{\beta}\beta + \frac{3}{2}\hat{\beta}(1 - \beta)] + a[\hat{\beta} + 12\hat{\beta}^2L\gamma + \hat{\beta}^2(1 - \beta)] \\
&\quad + \sqrt{2c\delta_{\text{byz}}}\hat{a}[8\hat{\beta}^2L\gamma + 12\hat{\beta}^2L\gamma + 4\hat{\beta}^2(1 - \beta)] \\
&\quad + \tau T\sqrt{2c\delta_{\text{byz}}}[3\frac{1}{2}\hat{\beta}^2(1 - \beta) + 3\hat{\beta}(1 - \hat{\beta}) + 12\hat{\beta}^2L\gamma + 8\hat{\beta}^2L\gamma]. \tag{26}
\end{aligned}$$

For the first coefficient, we have

$$\begin{aligned}
&2\hat{\beta} + 32\hat{\beta}L\gamma + 8(1 - \hat{\beta}) + 2\hat{\beta}(1 - \beta) \leq 8 \\
&\Leftrightarrow 2\hat{\beta} + 32\hat{\beta}L\gamma + 2\hat{\beta}(1 - \beta) \leq 8\hat{\beta} \\
&\Leftrightarrow 8L\gamma \leq 1, \tag{27}
\end{aligned}$$

where the last inequality is satisfied by the choice of the step-size $\gamma \leq \frac{1}{24L}$. For the second coefficient, we have

$$\begin{aligned}
&12\hat{\beta}L\gamma + 3(1 - \hat{\beta}) + \hat{\beta} + \frac{3}{2}\hat{\beta}(1 - \beta) \leq 3 \\
&\Leftrightarrow 12\hat{\beta}L\gamma + \hat{\beta} + \frac{3}{2}\hat{\beta}(1 - \beta) \leq 3\hat{\beta} \\
&\Leftrightarrow 24L\gamma \leq 1, \tag{28}
\end{aligned}$$

where the last inequality holds by the choice of the step-size $\gamma \leq \frac{1}{24L}$. For the third coefficient, we have

$$\begin{aligned}
&12\hat{\beta}L\gamma + 3(1 - \hat{\beta}) + \hat{\beta}\beta + \frac{3}{2}\hat{\beta}(1 - \beta) \leq 3 \\
&\Leftrightarrow 12\hat{\beta}L\gamma + \hat{\beta}\beta + \frac{3}{2}\hat{\beta}(1 - \beta) \leq 3\hat{\beta} \\
&\Leftrightarrow 24L\gamma \leq 1, \tag{29}
\end{aligned}$$

where the last inequality holds by the choice of the step-size $\gamma \leq \frac{1}{24L}$. For the fourth coefficient, we have

$$\hat{\beta} + 12\hat{\beta}^2L\gamma + \hat{\beta}^2(1 - \beta) \leq 3\hat{\beta} \Leftrightarrow 12L\gamma\hat{\beta}^2 + \hat{\beta}^2 \leq 2\hat{\beta} \Leftrightarrow 12L\gamma \leq 1, \tag{30}$$

where the second last inequality holds by the choice of the step-size $\gamma \leq \frac{1}{24L}$, and the last inequality holds by the choice of the momentum parameter $\hat{\beta} \leq 1$. For the fifth coefficient, we have

$$8\hat{\beta}^2L\gamma + 12\hat{\beta}^2L\gamma + 4\hat{\beta}^2(1 - \beta) \leq \hat{\beta} \Leftrightarrow 20\hat{\beta}L\gamma + 4\hat{\beta} \leq 1, \tag{31}$$

where the last inequality holds by the choice of the step-size $\gamma \leq \frac{1}{24L}$ and momentum parameter $\hat{\beta} \leq \frac{1}{24}$. For the sixth coefficient, we have

$$\begin{aligned}
& 3/2\hat{\beta}^2(1 - \beta) + 3\hat{\beta}(1 - \hat{\beta}) + 12\hat{\beta}^2L\gamma + 8\hat{\beta}^2L\gamma \leq 3 \\
& \Leftrightarrow 3/2\hat{\beta}^2 + 20\hat{\beta}^2L\gamma \leq 3\hat{\beta} \\
& \Leftrightarrow 20\hat{\beta}L\gamma \leq \frac{3}{2} \\
& \Leftrightarrow \frac{40}{3}\hat{\beta}L\gamma \leq 1,
\end{aligned} \tag{32}$$

where the last inequality holds by the choice of the step-size $\gamma \leq \frac{1}{24L}$ and momentum parameter $\hat{\beta} \leq 1$. $\square$

Lemma C.4. *Let each f_i be L -smooth, $\Delta \geq \Phi^0$, $B_{\text{init}} > \tau$. Assume that the following inequalities hold for the iterates generated by *Byz-Clip21-SGD2M**

1. $\gamma \leq \frac{1}{24L}$;
2. $12L\gamma \leq \beta$;
3. $\beta, \hat{\beta} \in [0, 1]$;
4. $\|\bar{m}^{t-1}\| \leq \sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\hat{\beta}a + \hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T$,
5. $\|\bar{g}^{t-1}\| \leq \sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T$;
6. $\|\nabla f_i(x^{t-1}) - v_i^{t-1}\| \leq \sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T$ for all $i \in \mathcal{G}$;
7. $\|v_i^t - g_i^{t-1}\| \leq B_{\text{init}}$ for all $i \in \mathcal{G}$;
8. $\|\Omega^t\| \leq a$;
9. $\|\Omega_i^{t-1}\| \leq \hat{a}$ for all $i \in \mathcal{G}$;
10. $\|\theta_i^t\| \leq b$ for all $i \in \mathcal{G}$;
11. $\Phi^{t-1} \leq 2\Delta$.

Then we have

$$\|\nabla f_i(x^t) - v_i^t\| \leq \sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T. \tag{33}$$

Proof. We have

$$\begin{aligned}
\|\nabla f_i(x^t) - v_i^t\| &\stackrel{(i)}{=} \|\nabla f_i(x^t) - (1-\beta)v_i^{t-1} - \beta\nabla f_i(x^t, \xi_i^t)\| \\
&\stackrel{(ii)}{\leq} (1-\beta)\|\nabla f_i(x^t) - v_i^{t-1}\| + \beta\|\nabla f_i(x^t) - \nabla f_i(x^t, \xi_i^t)\| \\
&\stackrel{(iii)}{\leq} (1-\beta)\|\nabla f_i(x^t) - \nabla f_i(x^{t-1})\| + (1-\beta)\|\nabla f_i(x^{t-1}) - v_i^{t-1}\| + \beta\|\theta_i^t\| \\
&\stackrel{(iv)}{\leq} (1-\beta)L\gamma\|g^{t-1}\| + (1-\beta)\|\nabla f_i(x^{t-1}) - v_i^{t-1}\| + \beta\|\theta_i^t\| \\
&\stackrel{(v)}{\leq} (1-\beta)L\gamma\|\bar{m}^{t-1} - g^{t-1}\| + (1-\beta)L\gamma\|\bar{m}^{t-1}\| + (1-\beta)\|\nabla f_i(x^{t-1}) - v_i^{t-1}\| + \beta\|\theta_i^t\| \\
&\stackrel{(vi)}{\leq} (1-\beta)L\gamma\left(2\hat{\beta}\tau T\sqrt{2c\delta_{\text{byz}}} + 2\hat{\beta}\sqrt{\frac{2c\delta_{\text{byz}}}{G}\sum_{i\in\mathcal{G}}\|\Omega_i^{t-1}\|^2}\right) + (1-\beta)L\gamma\|\bar{m}^{t-1}\| \\
&\quad + \beta\|\theta_i^t\| + (1-\beta)\|\nabla f_i(x^{t-1}) - v_i^{t-1}\| \\
&\stackrel{(vii)}{\leq} 2\hat{\beta}L\gamma\sqrt{2c\delta_{\text{byz}}}\tau T + 2\hat{\beta}L\gamma(1-\beta)\sqrt{2c\delta_{\text{byz}}}\hat{a} + \beta b \\
&\quad + (1-\beta)L\gamma\left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\hat{\beta}a + \hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T\right) \\
&\quad + (1-\beta)\left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + 3/2\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T\right) \\
&\leq \sqrt{L\Delta}[8L\gamma + 2(1-\beta)] + (B_{\text{init}} - \tau)[3L\gamma + 3^{(1-\beta)/2}] + b[3L\gamma + \beta + 3^{(1-\beta)/2}] \\
&\quad + \hat{\beta}a[3L\gamma + (1-\beta)] + \sqrt{2c\delta_{\text{byz}}}\hat{\beta}\hat{a}[2L\gamma + L\gamma + 4(1-\beta)] \\
&\quad + \sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T[2L\gamma + 3L\gamma + \frac{3}{2}(1-\beta)],
\end{aligned}$$

where (i) follows from the update rule of v_i^t , (ii)-(iii) – from triangle inequality and definition of θ_i^t , (iv) – from L -smoothness and the update rule of x^t , (v) – from triangle inequality (vi) – from (18) and inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$, (vii) – from the assumptions 4-7, 9, 10 of the lemma.

For the first coefficient, we have

$$8L\gamma + 2(1-\beta) \leq 2 \Leftarrow 4L\gamma \leq \beta \quad (34)$$

where the last inequality holds by the choice of the step-size $12L\gamma \leq \beta$. For the second coefficient, we have

$$3L\gamma + \frac{3}{2}(1-\beta) \leq \frac{3}{2} \Leftarrow 2L\gamma \leq \beta, \quad (35)$$

where the last inequality holds by the choice of the step-size $12L\gamma \leq \beta$. For the third coefficient, we have

$$3L\gamma + \beta + \frac{3}{2}(1-\beta) \leq \frac{3}{2} \Leftarrow 6L\gamma \leq \beta, \quad (36)$$

where the last inequality holds by the choice of the step-size $12L\gamma \leq \beta$. For the fourth coefficient, we have

$$3L\gamma + (1-\beta) \leq 1 \Leftarrow 3L\gamma \leq \beta, \quad (37)$$

the last inequality holds by the choice of the step-size $12L\gamma \leq \beta$. For the fifth coefficient, we have

$$2L\gamma + L\gamma + 4(1-\beta) \leq 4 \Leftarrow 3L\gamma \leq 4\beta, \quad (38)$$

where the last inequality holds by the choice of the step-size $12L\gamma \leq \beta$. For the sixth coefficient, we have

$$2L\gamma + 3L\gamma + \frac{3}{2}(1-\beta) \leq \frac{3}{2} \Leftarrow \frac{10}{3}L\gamma \leq \beta, \quad (39)$$

where the last inequality holds by the choice of the step-size $12L\gamma \leq \beta$. This concludes the proof. $\square$

Lemma C.5. *Let each f_i be L -smooth, $\Delta \geq \Phi^0$, $B_{\text{init}} > \tau$. Assume that the following inequalities hold for the iterates generated by *Byz-Clip21-SGD2M**

1. $\gamma \leq \frac{1}{24L}$;
2. $\beta \in [0, 1]$;
3. $\hat{\beta} \leq \min \left\{ \frac{\sqrt{L\Delta}}{a}, \frac{\sqrt{L\Delta}}{4\sqrt{2c\delta_{\text{byz}}}\hat{a}}, 1 \right\}$;
4. $\|\bar{m}^{t-1}\| \leq \sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\hat{\beta}a + \hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T$,
5. $\|\bar{g}^{t-1}\| \leq \sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T$;
6. $\|\nabla f_i(x^{t-1}) - v_i^{t-1}\| \leq \sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T$ for all $i \in \mathcal{G}$;
7. $\|v_i^t - g_i^{t-1}\| \leq B_{\text{init}}$ for all $i \in \mathcal{G}$;
8. $\|\Omega^t\| \leq a$;
9. $\|\Omega_i^{t-1}\| \leq \hat{a}$ for all $i \in \mathcal{G}$;
10. $\|\theta_i^t\| \leq b$ for all $i \in \mathcal{G}$;
11. $\Phi^{t-1} \leq 2\Delta$.

Then we have

$$\|\bar{g}^t\| \leq \sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T. \quad (40)$$

Proof. We have

$$\begin{aligned}
\|\bar{g}^t\| &\stackrel{(i)}{=} \left\| \bar{g}^{t-1} + \frac{1}{G} \sum_{i \in \mathcal{G}} \hat{\beta} \text{clip}_\tau(v_i^t - g_i^{t-1}) \right\| \\
&= \left\| \hat{\beta} \nabla f(x^t) + \hat{\beta}(\bar{v}^t - \nabla f(x^t)) + (1 - \hat{\beta})\bar{g}^{t-1} + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} [\text{clip}_\tau(v_i^t - g_i^{t-1}) - (v_i^t - g_i^{t-1})] \right\| \\
&\stackrel{(ii)}{\leq} \hat{\beta} \|\nabla f(x^t)\| + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} \|v_i^t - \nabla f_i(x^t)\| + (1 - \hat{\beta})\|\bar{g}^{t-1}\| \\
&\quad + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} \|\text{clip}_\tau(v_i^t - g_i^{t-1}) - (v_i^t - g_i^{t-1})\| \\
&\stackrel{(iii)}{\leq} \hat{\beta} \|\nabla f(x^{t-1})\| + \hat{\beta} \|\nabla f(x^t) - \nabla f(x^{t-1})\| + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} \|(1 - \beta)v_i^{t-1} + \beta \nabla f_i(x^t, \xi_i^t) - \nabla f_i(x^t)\| \\
&\quad + (1 - \hat{\beta})\|\bar{g}^{t-1}\| + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} \max \{ \|v_i^t - g_i^{t-1}\| - \tau, 0 \}, \quad (41)
\end{aligned}$$

where (i) follows from the update rule of g_i^t , (ii) – from the triangle inequality, (iii) – from the update rule of v_i^t , triangle

inequality, and Lemma B.1. We continue the derivation of the bound as follows

$$\begin{aligned}
\|\bar{g}^t\| &\stackrel{(iv)}{\leq} \hat{\beta}\sqrt{2L(f(x^{t-1}) - f^*)} + \hat{\beta}L\gamma\|g^{t-1}\| + (1 - \hat{\beta})\|\bar{g}^{t-1}\| + \hat{\beta}(B_{\text{init}} - \tau) \\
&\quad + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} ((1 - \beta)\|v_i^{t-1} - \nabla f_i(x^{t-1})\| + (1 - \beta)\|\nabla f_i(x^{t-1}) - \nabla f_i(x^t)\| + \beta\|\theta_i^t\|), \\
&\stackrel{(v)}{\leq} \hat{\beta}\sqrt{4L\Delta} + \hat{\beta}L\gamma(2 - \beta)\|g^{t-1}\| + (1 - \hat{\beta})\|\bar{g}^{t-1}\| + \hat{\beta}(B_{\text{init}} - \tau) + \hat{\beta}\beta b \\
&\quad + \frac{\hat{\beta}}{G}(1 - \beta) \sum_{i \in \mathcal{G}} \|v_i^{t-1} - \nabla f_i(x^{t-1})\| \\
&\stackrel{(vi)}{\leq} \hat{\beta}\sqrt{4L\Delta} + \hat{\beta}L\gamma(2 - \beta)\|\bar{m}^{t-1} - g^{t-1}\| + \hat{\beta}L\gamma(2 - \beta)\|\bar{m}^{t-1}\| + (1 - \hat{\beta})\|\bar{g}^{t-1}\| + \hat{\beta}(B_{\text{init}} - \tau) \\
&\quad + \hat{\beta}\beta b + \frac{\hat{\beta}}{G}(1 - \beta) \sum_{i \in \mathcal{G}} \|v_i^{t-1} - \nabla f_i(x^{t-1})\|, \tag{42}
\end{aligned}$$

where (iv) follows from L -smoothness, the update rule of x^t , assumption 7 of the lemma, and triangle inequality, (v) – from assumptions 10 and 11 of the lemma, L -smoothness, the update rule of x^t , (vi) – from triangle inequality. Using (18) we continue

$$\begin{aligned}
\|\bar{g}^t\| &\stackrel{(vii)}{\leq} \hat{\beta}\sqrt{4L\Delta} + 2L\gamma\hat{\beta} \left(2\hat{\beta}\tau T\sqrt{2c\delta_{\text{byz}}} + 2\hat{\beta}\sqrt{\frac{2c\delta_{\text{byz}}}{G} \sum_{i \in \mathcal{G}} \|\Omega_i^{t-1}\|^2} \right) + 2L\gamma\hat{\beta}\|\bar{m}^{t-1}\| \\
&\quad + (1 - \hat{\beta})\|\bar{g}^{t-1}\| + \hat{\beta}(B_{\text{init}} - \tau) + \hat{\beta}\beta b + \frac{\hat{\beta}}{G}(1 - \beta) \sum_{i \in \mathcal{G}} \|v_i^{t-1} - \nabla f_i(x^{t-1})\|, \tag{43}
\end{aligned}$$

where (vii) follows from (18) and inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any positive a, b . Then, we get

$$\begin{aligned}
\|\bar{g}^t\| &\stackrel{(viii)}{\leq} \hat{\beta}\sqrt{4L\Delta} + 4L\gamma\hat{\beta}^2\tau T\sqrt{2c\delta_{\text{byz}}} + 4L\gamma\hat{\beta}^2\sqrt{2c\delta_{\text{byz}}}\hat{a} + \hat{\beta}(B_{\text{init}} - \tau) + \hat{\beta}\beta b \\
&\quad + 4L\gamma\hat{\beta} \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\hat{\beta}a + \hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right) \\
&\quad + (1 - \hat{\beta}) \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right) \\
&\quad + \hat{\beta}(1 - \beta) \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right), \tag{44}
\end{aligned}$$

where (viii) follows from the assumptions 4-6 and 9 of the lemma. Regrouping the terms, we derive

$$\begin{aligned}
\|\bar{g}^t\| &\leq \sqrt{L\Delta}[2\hat{\beta} + 32L\gamma\hat{\beta} + 8(1 - \hat{\beta}) + 2\hat{\beta}] \\
&\quad + (B_{\text{init}} - \tau)[\hat{\beta} + 12L\gamma\hat{\beta} + 3(1 - \hat{\beta}) + 3\hat{\beta}(1-\beta)/2] \\
&\quad + b[\hat{\beta}\beta + 12L\gamma\hat{\beta} + 3(1 - \hat{\beta}) + 3\hat{\beta}(1-\beta)/2] + \hat{\beta}a[12L\gamma\hat{\beta} + (1 - \beta)\hat{\beta}] \\
&\quad + \hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a}[4L\gamma\hat{\beta} + 4L\gamma\hat{\beta} + 4\hat{\beta}(1 - \hat{\beta})] \\
&\quad + \sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T[4L\gamma\hat{\beta} + 12L\gamma\hat{\beta} + 3(1 - \hat{\beta}) + 3\hat{\beta}(1-\beta)/2], \tag{45}
\end{aligned}$$

For the second term, we have

$$\hat{\beta} + 12L\gamma\hat{\beta} + 3(1 - \hat{\beta}) + \frac{3}{2}\hat{\beta}(1 - \beta) \leq 3 \Leftarrow \hat{\beta} + 12L\gamma\hat{\beta} + \frac{3}{2}\hat{\beta} \leq 3\hat{\beta} \Leftarrow 24L\gamma \leq 1, \tag{46}$$

where the last inequality holds by the choice of the step-size $24L\gamma \leq 1$. For the third term, we have

$$\hat{\beta}\beta + 12L\gamma\hat{\beta} + 3(1 - \hat{\beta}) + \frac{3}{2}\hat{\beta}(1 - \beta) \leq 3 \Leftarrow \beta\hat{\beta} + 12L\gamma\hat{\beta} + \frac{3}{2}\hat{\beta} \leq 3\hat{\beta} \Leftarrow 24L\gamma \leq 1, \tag{47}$$

where the last inequality holds by the choice of the step-size $24L\gamma \leq 1$. For the fourth term, we have

$$\hat{\beta}a[12L\gamma\hat{\beta} + (1 - \beta)\hat{\beta}] \leq \sqrt{L\Delta}[12L\gamma\hat{\beta} + (1 - \beta)\hat{\beta}], \tag{48}$$

where we use $\hat{\beta} \leq \frac{\sqrt{L\Delta}}{a}$. Next, for the fifth term, we have

$$\hat{\beta}\sqrt{2c\delta_{\text{byz}}\hat{a}}[4L\gamma\hat{\beta} + 4L\gamma\hat{\beta} + 4\hat{\beta}(1 - \hat{\beta})] \leq \sqrt{L\Delta}[2L\gamma\hat{\beta} + \hat{\beta}(1 - \beta)], \quad (49)$$

where we use $\hat{\beta} \leq \frac{\sqrt{L\Delta}}{4\sqrt{2c\delta_{\text{byz}}\hat{a}}}$. Therefore, combining the two inequalities (48) and (49) with the first term, we obtain for the coefficient next to $\sqrt{L\Delta}$

$$\begin{aligned} & 12L\gamma\hat{\beta} + \hat{\beta}(1 - \beta) + 2L\gamma\hat{\beta} + \hat{\beta}(1 - \beta) + 2\hat{\beta} + 32L\gamma\hat{\beta} + 8(1 - \hat{\beta}) + 2\hat{\beta} \leq 8 \\ & \Leftrightarrow 46L\gamma\hat{\beta} + 2\hat{\beta}(1 - \beta) + 4\hat{\beta} \leq 8\hat{\beta} \\ & \Leftrightarrow 23L\gamma \leq 1, \end{aligned} \quad (50)$$

where the last inequality holds by the choice of the step-size $24L\gamma \leq 1$. For the sixth coefficient, we have

$$\begin{aligned} & 4L\gamma\hat{\beta} + 12L\gamma\hat{\beta} + 3(1 - \hat{\beta}) + \frac{3\hat{\beta}(1 - \beta)}{2} \leq 3 \Leftrightarrow 16L\gamma\hat{\beta} + \frac{3\hat{\beta}(1 - \beta)}{2} \leq 3\hat{\beta}, \\ & \Leftrightarrow \frac{32}{3}L\gamma \leq 1, \end{aligned} \quad (51)$$

where the last inequality holds by the choice of the step-size $24L\gamma \leq 1$. This concludes the proof. $\square$

Lemma C.6. *Let each f_i be L -smooth, $\Delta \geq \Phi^0$, $B_{\text{init}} > \tau$, and $i \in \mathcal{I}_t := \{i \in \mathcal{G} \mid \|v_i^t - g_i^{t-1}\| > \tau\}$. Assume that the following inequalities hold for the iterates generated by *Byz-Clip21-SGD2M*:*

1. $\gamma \leq \frac{1}{24L}$;
2. $\beta \leq \min \left\{ \frac{\hat{\beta}\tau}{54\sqrt{L\Delta}}, 1 \right\}$;
3. $\beta \leq \min \left\{ \frac{2\hat{\beta}\tau}{39(B_{\text{init}} - \tau)}, 1 \right\}$;
4. $\beta \leq \min \left\{ \frac{2\hat{\beta}\tau}{69b}, 1 \right\}$;
5. $\beta \leq \min \left\{ \frac{2}{41\sqrt{2c\delta_{\text{byz}}T}}, 1 \right\}$;
6. $\hat{\beta} \in \min \left\{ \frac{\sqrt{L\Delta}}{a}, \frac{\sqrt{L\Delta}}{4\sqrt{2c\delta_{\text{byz}}\hat{a}}}, \frac{1}{24} \right\}$;
7. $\|\bar{m}^{t-1}\| \leq \sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\hat{\beta}a + \hat{\beta}\sqrt{2c\delta_{\text{byz}}\hat{a}} + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T$,
8. $\|\bar{g}^{t-1}\| \leq \sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T$;
9. $\|\nabla f_i(x^{t-1}) - v_i^{t-1}\| \leq \sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}\hat{a}} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T$ for all $i \in \mathcal{G}$;
10. $\|\Omega_i^t\| \leq \hat{a}$ for all $i \in \mathcal{G}$;
11. $\|\theta_i^t\| \leq b$ and $\|\theta_i^{t+1}\| \leq b$ for all $i \in \mathcal{G}$.

Then we have

$$\|v_i^{t+1} - g_i^t\| \leq \|v_i^t - g_i^{t-1}\| - \frac{\tau\hat{\beta}}{2}. \quad (52)$$

Proof. Since $i \in \mathcal{I}_t$, then $\|v_i^t - g_i^{t-1}\| > \tau$ and from Lemma C.2 we have

$$\begin{aligned} \|v_i^{t+1} - g_i^t\| & \leq (1 - \hat{\beta})\|v_i^t - g_i^{t-1}\| + \hat{\beta}\|v_i^t - g_i^{t-1}\| - \hat{\beta}\tau + \beta L\gamma\|\bar{m}^t\| + \beta\|\nabla f_i(x^t) - v_i^t\| \\ & \quad + 2\hat{\beta}\beta L\gamma\tau T\sqrt{2c\delta_{\text{byz}}} + 2\hat{\beta}\beta L\gamma\sqrt{\frac{2c\delta_{\text{byz}}}{G}} \sum_{i \in \mathcal{G}} \|\Omega_i^t\|^2 + \beta\|\theta_i^{t+1}\|. \end{aligned} \quad (53)$$

Using the assumptions of the lemma and Lemmas C.3 and C.4, we obtain from (53)

$$\begin{aligned} \|v_i^{t+1} - g_i^t\| &\leq \|v_i^t - g_i^{t-1}\| - \hat{\beta}\tau + \beta L\gamma \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\hat{\beta}a + \hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right) \\ &\quad + \beta \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right) \\ &\quad + 2\hat{\beta}\beta L\gamma\tau T\sqrt{2c\delta_{\text{byz}}} + 2\beta\hat{\beta}L\gamma\sqrt{2c\delta_{\text{byz}}}\hat{a} + \beta b. \end{aligned} \quad (54)$$

Regrouping the terms, we derive

$$\begin{aligned} \|v_i^{t+1} - g_i^t\| &\leq \|v_i^t - g_i^{t-1}\| - \hat{\beta}\tau + \sqrt{L\Delta}[8L\gamma\beta + 2\beta] + (B_{\text{init}} - \tau)[3L\gamma\beta + 3\beta/2] \\ &\quad + b[3L\gamma\beta + 3\beta/2 + \beta] + \hat{\beta}a[3L\gamma\beta + \beta] + \hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a}[3\beta L\gamma + 4\beta] \\ &\quad + \hat{\beta}\tau T\sqrt{2c\delta_{\text{byz}}}[3\beta L\gamma + 3\beta/2 + 2\beta L\gamma] \\ &= \|v_i^t - g_i^{t-1}\| - \hat{\beta}\tau + \sqrt{L\Delta}[8L\gamma\beta + 2\beta] + (B_{\text{init}} - \tau)[3L\gamma\beta + 3\beta/2] \\ &\quad + b[3L\gamma\beta + 5\beta/2] + \hat{\beta}a[3L\gamma\beta + \beta] + \hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a}[3\beta L\gamma + 4\beta] \\ &\quad + \hat{\beta}\tau T\sqrt{2c\delta_{\text{byz}}}[5\beta L\gamma + 3\beta/2] \end{aligned} \quad (55)$$

First, note that using restrictions $\hat{\beta}a \leq \sqrt{L\Delta}$, we obtain

$$\hat{\beta}a[3L\gamma\beta + \beta] \leq \sqrt{L\Delta}[3L\gamma\beta + \beta]. \quad (56)$$

Next, we also have

$$\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a}[3\beta L\gamma + 4\beta] \leq \frac{\sqrt{L\Delta}}{4}[3\beta L\gamma + 4\beta] = \sqrt{L\Delta}[3/4\beta L\gamma + \beta]. \quad (57)$$

Combining (56) and (57) with other terms involving $\sqrt{L\Delta}$ and using $24L\gamma \leq 1$, we have

$$\sqrt{L\Delta}[3L\gamma\beta + \beta + 3/4\beta L\gamma + \beta + 8L\gamma\beta + 2\beta] \leq \sqrt{L\Delta}[12L\gamma\beta + 4\beta] \leq \sqrt{L\Delta}\frac{9\beta}{2} \leq \frac{\hat{\beta}\tau}{12}, \quad (58)$$

where we used $\beta \leq \frac{\hat{\beta}\tau}{54\sqrt{L\Delta}}$ and $24L\gamma \leq 1$. Again, since $24L\gamma \leq 1$, we have

$$[3L\gamma\beta + 3\beta/2](B_{\text{init}} - \tau) \leq \frac{13\beta}{8}(B_{\text{init}} - \tau) \leq \frac{\hat{\beta}\tau}{12}, \quad (59)$$

where we used $\beta \leq \frac{2\hat{\beta}\tau}{39(B_{\text{init}} - \tau)}$. Since $24L\gamma \leq 1$, we have

$$[3L\gamma\beta + 3\beta/2 + \beta]b \leq \frac{23\beta}{8}b \leq \frac{\hat{\beta}\tau}{12}, \quad (60)$$

where we use $\beta \leq \frac{2\hat{\beta}\tau}{69b}$. Since $24L\gamma \leq 1$, we have

$$[5\beta L\gamma + 3\beta/2]\hat{\beta}\tau T\sqrt{2c\delta_{\text{byz}}} \leq \frac{41}{24}\beta\hat{\beta}\tau T\sqrt{2c\delta_{\text{byz}}} \leq \frac{\hat{\beta}\tau}{12}, \quad (61)$$

where we use $\beta \leq \frac{2}{41\sqrt{2c\delta_{\text{byz}}T}}$. Combining all the bounds, we obtain from (55) that

$$\begin{aligned} \|v_i^{t+1} - g_i^t\| &\leq \|v_i^t - g_i^{t-1}\| - \hat{\beta}\tau + \sqrt{L\Delta}[8L\gamma\beta + 2\beta] + (B_{\text{init}} - \tau)[3L\gamma\beta + 3\beta/2] \\ &\quad + b[3L\gamma\beta + 5\beta/2] + \hat{\beta}a[3L\gamma\beta + \beta] + \hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a}[3\beta L\gamma + 4\beta] \\ &\quad + \hat{\beta}\tau T\sqrt{2c\delta_{\text{byz}}}[5\beta L\gamma + 3\beta/2] \\ &\leq \|v_i^t - g_i^{t-1}\| - \hat{\beta}\tau + 4 \cdot \frac{\tau\hat{\beta}}{12} + \frac{\tau\hat{\beta}}{6} \\ &= \|v_i^t - g_i^{t-1}\| - \frac{\hat{\beta}\tau}{2}. \end{aligned} \quad (62)$$

□

Lemma C.7 (Lemma 17 from [Islamov et al. \[2025b\]](#)). *Let $\|\theta_i^{t+1}\| \leq b$ for all $i \in \mathcal{G}$. Let each f_i be L -smooth. Then, for the iterates generated by *Byz-Clip21-SGD2M* the quantity $\tilde{P}^t := \frac{1}{G} \sum_{i \in \mathcal{G}} \|v_i^t - \nabla f_i(x^t)\|^2$ decreases as*

$$\tilde{P}^{t+1} \leq (1 - \beta)\tilde{P}^t + \frac{3L^2}{\beta}R^t + \beta^2b^2 + \frac{2}{G}\beta(1 - \beta) \sum_{i \in \mathcal{G}} \langle v_i^t - \nabla f_i(x^{t+1}), \theta_i^{t+1} \rangle, \quad (63)$$

where $R^t := \|x^{t+1} - x^t\|^2$ and $\theta_i^t = \nabla f_i(x^t, \xi_i^t) - \nabla f_i(x^t)$.

Lemma C.8 (Lemma 18 from [Islamov et al. \[2025b\]](#)). *Let $\|\theta^{t+1}\| \leq \frac{\hat{c}}{\sqrt{G}}$. Let each f_i be L -smooth. Then, for the iterates generated by *Byz-Clip21-SGD2M* the quantity $\tilde{P}^t := \|\bar{v}^t - \nabla f(x^t)\|^2$ decreases as*

$$\tilde{P}^{t+1} \leq (1 - \beta)\tilde{P}^t + \frac{3L^2}{\beta}R^t + \beta^2\frac{\hat{c}^2}{G} + 2\beta(1 - \beta)\langle \bar{v}^t - \nabla f(x^{t+1}), \theta^{t+1} \rangle, \quad (64)$$

where $R^t := \|x^{t+1} - x^t\|^2$ and $\theta^t = \frac{1}{G} \sum_{i \in \mathcal{G}} \nabla f_i(x^t, \xi_i^t) - \nabla f_i(x^t)$.

Lemma C.9 (Lemma 19 from [Islamov et al. \[2025b\]](#)). *Let $\|\theta_i^t\| \leq b$ for all $i \in \mathcal{G}$, each f_i be L -smooth, and $\|v_i^t - g_i^{t-1}\| \leq B_{\text{init}}$ for all $i \in \mathcal{G}$ and some $B_{\text{init}} > \tau$, and $\hat{\beta} \leq \frac{1}{2\eta}$. Then for the iterates generated by *Byz-Clip21-SGD2M* we have*

$$\begin{aligned} \|g_i^t - v_i^t\|^2 &\leq (1 - \hat{\beta}\eta)\|g_i^{t-1} - v_i^{t-1}\|^2 + \frac{4\beta^2}{\hat{\beta}\eta}\|v_i^{t-1} - \nabla f_i(x^{t-1})\|^2 + \frac{4\beta^2L^2}{\hat{\beta}\eta}R^{t-1} + \beta^2b^2 \\ &\quad + 2(1 - \hat{\beta}\eta)^2\beta\langle (g_i^{t-1} - v_i^{t-1}) + \beta(v_i^{t-1} - \nabla f_i(x^{t-1})), \theta_i^t \rangle \\ &\quad + 2(1 - \hat{\beta}\eta)^2\beta^2\langle \nabla f_i(x^{t-1}) - \nabla f_i(x^t), \theta_i^t \rangle, \end{aligned} \quad (65)$$

where $R^t := \|x^{t+1} - x^t\|^2$ and $\eta := \frac{\tau}{B_{\text{init}}} \in (0, 1)$. Moreover, averaging the inequalities over $i \in \mathcal{G}$, we get

$$\begin{aligned} \tilde{V}^t &\leq (1 - \hat{\beta}\eta)\tilde{V}^{t-1} + \frac{4\beta^2}{\hat{\beta}\eta}\tilde{P}^{t-1} + \frac{4\beta^2L^2}{\hat{\beta}\eta}R^{t-1} + \beta^2b^2 \\ &\quad + \frac{2}{G}(1 - \hat{\beta}\eta)^2\beta \sum_{i \in \mathcal{G}} \langle (g_i^{t-1} - v_i^{t-1}) + \beta(v_i^{t-1} - \nabla f_i(x^{t-1})) + \beta(\nabla f_i(x^{t-1}) - \nabla f_i(x^t)), \theta_i^t \rangle, \end{aligned} \quad (66)$$

where $\tilde{V}^t := \frac{1}{G} \sum_{i \in \mathcal{G}} \|g_i^t - v_i^t\|^2$ and $\tilde{P}^t := \frac{1}{G} \sum_{i \in \mathcal{G}} \|v_i^t - \nabla f_i(x^t)\|^2$.

Let us define constants that we will use in the proof of the main theorem.

$$\begin{aligned} a &:= \left(\sqrt{2} + 2\sqrt{3 \log \frac{8(T+1)}{\alpha}} \right) \sqrt{d}\sigma_\omega \sqrt{\frac{T}{G}}, \\ \hat{a} &:= \left(\sqrt{2} + 2\sqrt{3 \log \frac{8G(T+1)}{\alpha}} \right) \sqrt{d}\sigma_\omega \sqrt{T}, \\ b^2 &:= 2\sigma^2 \log \left(\frac{16(T+1)G}{\alpha} \right), \\ \hat{c}^2 &:= \left(\sqrt{2} + 2\sqrt{3 \log \frac{8(T+1)}{\alpha}} \right)^2 \sigma^2, \end{aligned} \quad (67)$$

D PROOF OF THEOREM 5.1

Theorem D.1 (Full statement of Theorem 5.1). *Let Assumptions 3.3 and 3.4, and define $B_{\text{init}} := \max\{3\tau, \max_i\{\|\nabla f_i(x^0)\|\} + b\}$. Let the failure probability α be such that $\alpha \in (0, 1)$, and constants $a, \hat{a}, b$, and $\hat{c}$ be defined as in (67), and $\Delta \geq \Phi^0$ for Φ^0 defined in (9). Consider the run of *Byz-Clip21-SGD2M* (Algorithm 1) for T iterations with DP noise variance σ_ω . Assume the following inequalities hold*

1. step-size restrictions:

- (a) $\gamma \leq \frac{1}{24L}$;
- (b) $12L\gamma = \beta$;
- (c)

$$\frac{5}{6} - \frac{32L^2\beta^2}{\hat{\beta}^2\eta^2}\gamma^2 - \frac{96L^2}{\hat{\beta}^2\eta^2}\gamma^2 \geq 0. \quad (68)$$

2. momentum restrictions:

- (a) $\beta \leq \min \left\{ \frac{\hat{\beta}\tau}{54\sqrt{L\Delta}}, 1 \right\}$;
- (b) $\beta \leq \min \left\{ \frac{2\hat{\beta}\tau}{39(B_{\text{init}} - \tau)}, 1 \right\}$;
- (c) $\beta \leq \min \left\{ \frac{2\hat{\beta}\tau}{69b}, 1 \right\}$;
- (d) $\beta \leq \min \left\{ \frac{2}{41\sqrt{2c\delta_{\text{byz}}T}}, 1 \right\}$;
- (e) $\hat{\beta} \in \min \left\{ \frac{\sqrt{L\Delta}}{a}, \frac{\sqrt{L\Delta}}{4\sqrt{2c\delta_{\text{byz}}\hat{a}}}, 1 \right\}$;
- (f) and momentum restrictions defined in (71), (72), (73), (74), (75), (77), (76), and (78);

Then, with probability $1 - \alpha$, we bound $\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|^2$ with

$$\sqrt{L\Delta}(\sqrt{L\Delta} + B_{\text{init}} + \sigma)\tilde{\mathcal{O}} \left(\frac{\sqrt{d}\sigma_\omega}{\sqrt{GT}\tau} + \frac{d^{1/3}\sigma_\omega^{2/3}}{(TG)^{1/3}\tau^{2/3}} + \frac{\sqrt{c\delta_{\text{byz}}}\sqrt{d}\sigma_\omega}{\sqrt{T}\tau} + \sqrt{c\delta_{\text{byz}}} \right),$$

where $\tilde{\mathcal{O}}$ hides constant and logarithmic factors and higher order terms decreasing in T . Let additionally Assumption 3.5a hold, then

$$\sqrt{L\Delta}(\sqrt{L\Delta} + B_{\text{init}} + \sigma) = \tilde{\mathcal{O}} \left((1 + B_\zeta)LF^0 + \zeta^2 + \sigma((1 + \sqrt{B_\zeta})\sqrt{LF^0} + \zeta) \right).$$

Proof. For convenience, we define $\nabla f_i(x^{-1}, \xi_i^{-1}) = v_i^{-1} = g_i^{-1} = 0, \Phi^{-1} = \Phi^0$. Next, let us define an event E^t for each $t \in \{0, \dots, T\}$ such that the following inequalities hold for all $k \in \{0, \dots, t\}$

1. $\|v_i^k - g_i^{k-1}\| \leq B_{\text{init}}$ for $i \in \mathcal{I}_k$;
2. $\|\bar{m}^t\| \leq \sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\hat{\beta}a + \hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T$,
3. $\|\bar{g}^t\| \leq \sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T$;
4. $\|\nabla f_i(x^t) - v_i^t\| \leq \sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T$ for all $i \in \mathcal{G}$;
5. $\|\theta_i^k\| \leq b$ for all $i \in \mathcal{G}$ and $\|\theta^k\| \leq \frac{\hat{c}}{\sqrt{G}}$;
6. $\left\| \frac{1}{G} \sum_{l=1}^{k+1} \sum_{i \in \mathcal{G}} \omega_i^l \right\| \leq a$ and $\left\| \sum_{l=1}^{k+1} \omega_i^l \right\| \leq \hat{a}$ for all $i \in \mathcal{G}$;
7. $\Phi^k \leq 2\Delta$;
- 8.

$$\begin{aligned} \frac{7}{8}\Delta &\geq \frac{4\gamma\beta}{n\hat{\beta}\eta}(1-\eta)^2 \sum_{l=0}^{k-1} \sum_{i \in \mathcal{G}} \langle (g_i^l - v_i^l) + \beta(v_i^l - \nabla f_i(x^l)) + \beta(\nabla f_i(x^l) - \nabla f_i(x^{l+1})), \theta_i^t \rangle \\ &\quad + \frac{16\gamma\beta^2}{n\hat{\beta}^2\eta^2}(1-\beta) \sum_{l=0}^{k-1} \sum_{i \in \mathcal{G}} \langle v_i^l - \nabla f_i(x^l), \theta_i^{l+1} \rangle + 4\gamma(1-\beta) \sum_{l=0}^{k-1} \langle v^l - \nabla f(x^l), \theta^{l+1} \rangle \\ &\quad + \frac{15\gamma\beta^2}{n\hat{\beta}^2\eta^2}(1-\beta) \sum_{l=0}^{k-1} \sum_{i \in \mathcal{G}} \langle \nabla f_i(x^l) - \nabla f_i(x^{l+1}), \theta_i^{l+1} \rangle \\ &\quad + 4\gamma(1-\beta) \sum_{l=0}^{k-1} \langle \nabla f(x^l) - \nabla f(x^{l+1}), \theta^{l+1} \rangle. \end{aligned}$$

Then, we will derive the result by induction, i.e., using the induction w.r.t. t , we will show that $\mathbb{P}(E^t) \geq 1 - \frac{\alpha(t+1)}{T+1}$ for all $t \in \{0, \dots, T-1\}$.

Before moving on to the proof's induction part, we need to establish several useful bounds. Denote the events Θ_i^t, Θ^t and N^{t+1} as

$$\begin{aligned}\Theta_i^t &:= \{\|\theta_i^t\| \geq b\}, \quad \Theta^t := \left\{\|\theta^t\| \geq \frac{\hat{c}}{\sqrt{G}}\right\}, \quad N^{t+1} := \left\{\left\|\frac{1}{G} \sum_{l=1}^t \sum_{i \in \mathcal{G}} \omega_i^l\right\| \geq a\right\}, \quad \text{and} \\ N_i^{t+1} &= \left\{\left\|\sum_{l=1}^t \omega_i^l\right\| \geq \hat{a}\right\}\end{aligned}\tag{69}$$

respectively. From Assumption 3.4 we have

$$\mathbb{P}(\Theta_i^t) \leq 2 \exp\left(-\frac{b^2}{2\sigma^2}\right) = \frac{\alpha}{8(T+1)G}$$

where the last equality is by definition of b^2 . Therefore, $\mathbb{P}(\bar{\Theta}_i^t) \geq 1 - \frac{\alpha}{8(T+1)G}$. Besides, notice that the constant $\hat{c}$ in (67) can be viewed as

$$\hat{c}^2 = (\sqrt{2} + 2b_3)\sigma^2 \quad \text{where} \quad b_3^2 = 3 \log \frac{8(T+1)}{\alpha}.$$

Now, we can use Lemma B.2 to bound $\mathbb{P}(\Theta^t)$. Since all θ_i^t are independent σ -sub-Gaussian random vectors, then we have

$$\mathbb{P}\left(\left\|\sum_{i \in \mathcal{G}} \theta_i^t\right\| \geq \hat{c}\sqrt{G}\right) = \mathbb{P}\left(\|\theta^t\| \geq \frac{\hat{c}}{\sqrt{G}}\right) \leq \exp(-b_3^2/3) = \frac{\alpha}{8(T+1)}.$$

We also use Lemma B.2 to bound $\mathbb{P}(N^t)$ and $\mathbb{P}(N_i^t)$. Indeed, since all ω_i^l are independent Gaussian random vectors, then we have

$$\mathbb{P}\left(\left\|\sum_{l=1}^t \sum_{i \in \mathcal{G}} \omega_i^l\right\| \geq (\sqrt{2} + 2b_2)\sqrt{\sum_{l=1}^t \sum_{i \in \mathcal{G}} \sigma_\omega^2 d}\right) \leq \exp(-b_2^2/3) = \frac{\alpha}{8(T+1)}.$$

with $b_2^2 = 3 \log \left(\frac{8(T+1)}{\alpha}\right)$. This implies that

$$\mathbb{P}\left(\left\|\frac{1}{G} \sum_{l=1}^t \sum_{i \in \mathcal{G}} \omega_i^l\right\| \geq a\right) \leq \frac{\alpha}{8(T+1)}$$

due to the choice of a from (67):

$$a = (\sqrt{2} + 2b_2)\sigma_\omega \sqrt{d} \sqrt{\frac{T}{G}}, \quad \text{where} \quad b_2^2 = 3 \log \frac{8(T+1)}{\alpha}.$$

Note that with this choice of a we have that the above is true for any $t \in \{1, \dots, T\}$, i.e., $\mathbb{P}(N^t) \geq 1 - \frac{\alpha}{8(T+1)}$ for all $t \in \{1, \dots, T\}$. Similarly, we derive

$$\mathbb{P}\left(\left\|\sum_{l=1}^t \omega_i^l\right\| \geq (\sqrt{2} + 2b_4)\sqrt{\sum_{l=1}^t \sigma_\omega^2 d}\right) \leq \exp(-b_4^2/3) \leq \frac{\alpha}{8G(T+1)}$$

with $b_4^2 = 3 \log \left(\frac{8G(T+1)}{\alpha}\right)$. Again, this implies that

$$\mathbb{P}\left(\left\|\sum_{l=1}^t \omega_i^l\right\| \geq \hat{a}\right) \leq \frac{\alpha}{8G(T+1)}$$

due to the choice of $\hat{a}$ from (67)

$$\hat{a} = (\sqrt{2} + 2b_4)\sigma_\omega \sqrt{T}, \quad \text{where} \quad b_4^2 = 3 \log \left(\frac{8G(T+1)}{\alpha}\right)$$

Now, we are ready to prove that $\mathbb{P}(E^t) \geq 1 - \frac{\alpha(t+1)}{T+1}$ for all $t \in \{0, \dots, T-1\}$. First, we show that the base of induction holds.

Base of induction.

1. $\|v_i^0 - g_i^{-1}\| = \|v_i^0\| = \beta\|\nabla f_i(x^0, \xi_i^0)\| = \beta\|\theta_i^0\| + \beta\|\nabla f_i(x^0)\| \leq \frac{1}{2}b + \frac{1}{2}B_{\text{init}} \leq \frac{1}{2}B_{\text{init}} + \frac{1}{2}B_{\text{init}} = B_{\text{init}}$ holds with probability $1 - \frac{\alpha}{8(T+1)}$. Indeed, we have

$$\mathbb{P}(\Theta_i^0) \leq 2 \exp\left(-\frac{b^2}{2\sigma^2}\right) = \frac{\alpha}{8(T+1)G}.$$

Therefore, we have

$$\mathbb{P}\left(\cap_{i \in \mathcal{G}} \bar{\Theta}_i^0\right) = 1 - \mathbb{P}\left(\cup_{i \in \mathcal{G}} \Theta_i^0\right) \geq 1 - \sum_{i \in \mathcal{G}} \mathbb{P}(\Theta_i^0) = 1 - G \frac{\alpha}{8(T+1)G} = 1 - \frac{\alpha}{8(T+1)}.$$

Moreover, we have

$$\mathbb{P}(\Theta^0) \leq \frac{\alpha}{8(T+1)}.$$

This means that the probability of the event that each $\left\|\frac{1}{G} \sum_{l=1}^t \sum_{i \in \mathcal{G}} \omega_i^l\right\| \leq a$, $\|\sum_{l=1}^t \omega_i^l\| \leq \hat{a}$, $\|\theta_i^0\| \leq b$, and $\|\theta^0\| \leq \frac{c}{\sqrt{G}}$, and is at least

$$1 - \frac{\alpha}{8(T+1)} - n \frac{\alpha}{8G(T+1)} - \frac{\alpha}{8(T+1)} - n \frac{\alpha}{8G(T+1)} = 1 - \frac{\alpha}{2(T+1)}.$$

2. We have already shown that

$$\mathbb{P}\left(\left\|\frac{1}{G} \sum_{i \in \mathcal{G}} \omega_i^1\right\| \geq a\right) \leq \frac{\alpha}{8(T+1)},$$

implying that $\left\|\frac{1}{G} \sum_{i \in \mathcal{G}} \omega_i^1\right\| \leq a$ with probability at least $1 - \frac{\alpha}{8(T+1)}$. Similarly, we have shown that $\|\sum_{l=1}^1 \omega_i^l\| \leq \hat{a}$ with probability $1 - \frac{\alpha}{8G(T+1)}$. Therefore,

3. $\bar{g}^0 = \frac{1}{G} \sum_{i \in \mathcal{G}} (g_i^{-1} + \hat{\beta} \text{clip}_\tau(v_i^0 - g_i^{-1})) = \frac{1}{G} \sum_{i \in \mathcal{G}} \hat{\beta} \text{clip}_\tau(\beta \nabla f_i(x^0, \xi_i^0))$. Therefore, we have

$$\begin{aligned} \|\bar{g}^0\| &= \left\| \frac{1}{G} \sum_{i \in \mathcal{G}} \hat{\beta} \nabla f_i(x^0) + \hat{\beta} \beta \theta_i^0 + (\hat{\beta} \text{clip}_\tau(\beta \nabla f_i(x^0, \xi_i^0)) - \hat{\beta} \beta \nabla f_i(x^0, \xi_i^0)) \right\| \\ &\leq \hat{\beta} \beta \|\nabla f(x^0)\| + \frac{\hat{\beta} \beta}{G} \sum_{i \in \mathcal{G}} \|\theta_i^0\| + \frac{1}{G} \sum_{i \in \mathcal{G}} \max\{0, \beta \|\nabla f_i(x^0, \xi_i^0)\| - \tau\} \\ &\leq \hat{\beta} \beta \sqrt{2L(f(x^0) - f(x^*))} + \frac{\hat{\beta} \beta}{G} \sum_{i \in \mathcal{G}} \|\theta_i^0\| + \frac{\hat{\beta}}{G} \sum_{i \in \mathcal{G}} \max\{0, \beta \|\nabla f_i(x^0)\| + \beta \|\theta_i^0\| - \tau\}. \end{aligned}$$

Note that if one of the maximum terms in the last sum is actually zero, then the bound is easier to satisfy. Therefore, we consider only the case when the maximum is always equal to the second term. We continue as follows

$$\begin{aligned} \|\bar{g}^0\| &\leq \frac{1}{2} \sqrt{2L\Phi^0} + \frac{2\hat{\beta}\beta}{G} \sum_{i \in \mathcal{G}} \|\theta_i^0\| + \frac{\hat{\beta}\beta}{G} \sum_{i \in \mathcal{G}} \|\nabla f_i(x^0)\| - \hat{\beta}\tau \\ &\leq \sqrt{64L\Delta} + 2\hat{\beta}\beta b + \hat{\beta}\beta B_{\text{init}} - \hat{\beta}\tau \\ &\leq \sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T. \end{aligned}$$

The inequalities above again hold in $\cap_{i \in \mathcal{G}} \bar{\Theta}_i^0$, i.e., with probability at least $1 - \frac{\alpha}{8(T+1)}$. Therefore, the condition 3 of the induction is verified. Since at iteration 0 we have $\bar{m}^0 = \bar{g}^0$, the condition 2 holds as well.

4. We have

$$\begin{aligned} \|v_i^0 - \nabla f_i(x^0)\| &= \|\beta \nabla f_i(x^0, \xi_i^0) - \nabla f_i(x^0)\| \\ &\leq \beta \|\nabla f_i(x^0, \xi_i^0) - \nabla f_i(x^0)\| + (1 - \beta) \|\nabla f_i(x^0)\| \\ &\leq \beta b + (1 - \beta) B_{\text{init}} \end{aligned}$$

The bound above holds with probability at least $1 - \frac{\alpha}{8(T+1)}$ because it holds in $\cap_{i \in \mathcal{G}} \bar{\Theta}_i^0$. Therefore, the bound 5 of the assumption of the induction is verified.

5. Next, we emphasize that the condition 8 of the induction assumption also hold, as $\Phi^0 \leq 2\Phi^0 \leq 2\Delta$ by the choice of Δ .
6. We finalize the induction base by noting that the condition 9 of the induction assumption holds since the RHS equals 0.

Therefore, we conclude that the conditions 1-8 hold with a probability of at least

$$\begin{aligned} \mathbb{P}\left(\bar{\Theta}^0 \cap \left(\cap_{i \in \mathcal{G}} \bar{\Theta}_i^0\right) \cap \bar{N}^t \cap \left(\cap_{i \in \mathcal{G}} \bar{N}_i^t\right)\right) &\geq 1 - \mathbb{P}(\Theta^0) - \sum_{i \in \mathcal{G}} \mathbb{P}(\Theta_i^0) - \mathbb{P}(N^0) - \sum_{i \in \mathcal{G}} \mathbb{P}(N_i^0) \\ &\geq 1 - \frac{\alpha}{8(T+1)} - G \cdot \frac{\alpha}{8G(T+1)} - \frac{\alpha}{8(T+1)} - G \cdot \frac{\alpha}{8G(T+1)} \\ &= 1 - \frac{\alpha}{2(T+1)} > 1 - \frac{\alpha}{T+1}, \end{aligned}$$

i.e., $\mathbb{P}(E^0) \geq 1 - \frac{\alpha}{T+1}$ holds. This is the base of the induction.

Transition step of induction. **Case** $|\mathcal{I}_{K+1}| > 0$. Assume that all events $\bar{\Theta}^{K+1}$, $\bar{\Theta}_i^{K+1}$, $\bar{N}^{K+1}$, and $\bar{N}_i^{K+1}$ take place, i.e., $\|\theta_i^{K+1}\| \leq b$, $\|\theta^{K+1}\| \leq \frac{c}{\sqrt{G}}$ for all $i \in \mathcal{G}$, $\left\|\frac{1}{G} \sum_{l=1}^K \sum_{i \in \mathcal{G}} \omega_i^l\right\| \leq a$, and $\left\|\sum_{l=1}^K \omega_i^l\right\| \leq \hat{a}$ for all $i \in \mathcal{G}$. That is, we assume that the event

$$\bar{\Theta}^{K+1} \cap \left(\cap_{i \in \mathcal{G}} \bar{\Theta}_i^{K+1}\right) \cap \bar{N}^{K+1} \cap \left(\cap_{i \in \mathcal{G}} \bar{N}_i^{K+1}\right) \cap E^K$$

holds. Then, by the assumptions of the induction and from Lemma C.6 we get for all $i \in \mathcal{I}_{K+1}$

$$\|v_i^{K+1} - g_i^K\| \leq \|v_i^K - g_i^{K-1}\| - \frac{\hat{\beta}\tau}{2} \leq B_{\text{init}} - \frac{\hat{\beta}\tau}{2}.$$

Therefore, from Lemma C.3 we get that

$$\|\bar{m}^{K+1}\| \leq \sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\hat{\beta}a + \hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T,$$

from Lemma C.5 we get that

$$\|\bar{g}^{K+1}\| \leq \sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T,$$

from Lemma C.4

$$\|\nabla f_i(x^{K+1}) - v_i^{K+1}\| \leq \sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \quad \text{for all } i \in \mathcal{G}.$$

This means that conditions 1-7 in the induction assumption are also verified for the step $K+1$. Since for all $t \in \{0, \dots, K+1\}$ inequalities 1-7 are verified, we can write for each $t \in \{0, \dots, K\}$ by Lemmas C.1 and C.7 to C.9 the following

$$\begin{aligned} \Phi^{t+1} &= \delta^{t+1} + \frac{2\gamma}{\hat{\beta}\eta} \tilde{V}^{t+1} + \frac{8\gamma\beta}{\hat{\beta}^2\eta^2} \tilde{P}^{t+1} + \frac{2\gamma}{\beta} P^{t+1} \\ &\leq \delta^t - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \frac{1}{4\gamma} R^t + 2\gamma \tilde{V}^t + 2\gamma P^t + 2\gamma \hat{\beta}^2 \|\Omega^t\|^2 + 16\gamma \hat{\beta}^2 c\delta_{\text{byz}} \tau^2 T^2 + \frac{16\gamma \hat{\beta}^2 c\delta_{\text{byz}}}{G} \sum_{i \in \mathcal{G}} \|\Omega_i^t\|^2 \\ &\quad + \frac{2\gamma}{\hat{\beta}\eta} \left((1 - \hat{\beta}\eta) \tilde{V}^t + \frac{4\beta^2}{\hat{\beta}\eta} \tilde{P}^t + \frac{4\beta^2 L^2}{\hat{\beta}\eta} R^t + \beta^2 b^2 \right. \\ &\quad \left. + \frac{2}{G} \beta (1 - \hat{\beta}\eta)^2 \sum_{i \in \mathcal{G}} \langle (g_i^t - v_i^t) + \beta(v_i^t - \nabla f_i(x^t)) + \beta(\nabla f_i(x^t) - \nabla f_i(x^{t+1})), \theta_i^{t+1} \rangle \right) \\ &\quad + \frac{8\gamma\beta}{\hat{\beta}^2\eta^2} \left((1 - \beta) \tilde{P}^t + \frac{3L^2}{\beta} R^t + \beta^2 b^2 + \frac{2}{G} \beta (1 - \beta) \sum_{i \in \mathcal{G}} \langle v_i^t - \nabla f_i(x^{t+1}), \theta_i^{t+1} \rangle \right) \\ &\quad + \frac{2\gamma}{\beta} \left((1 - \beta) P^t + \frac{3L^2}{\beta} R^t + \beta^2 \frac{\hat{c}^2}{G} + 2\beta(1 - \beta) \langle v^t - \nabla f(x^{t+1}), \theta^{t+1} \rangle \right) \end{aligned}$$

Rearranging terms, we get

$$\begin{aligned}
\Phi^{t+1} \leq & \delta^t - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 + \frac{2\gamma}{\hat{\beta}\eta} \tilde{V}^t (\hat{\beta}\eta + 1 - \hat{\beta}\eta) + \frac{8\gamma\beta}{\hat{\beta}^2\eta^2} \tilde{P}^t (\beta + 1 - \beta) + \frac{2\gamma}{\beta} P^t (\beta + 1 - \beta) \\
& + 2\gamma\hat{\beta}^2 \|\Omega^t\|^2 + 16\gamma\hat{\beta}^2 c\delta_{\text{byz}} \tau^2 T^2 + 16\gamma\hat{\beta}^2 \frac{c\delta_{\text{byz}}}{G} \sum_{i \in \mathcal{G}} \|\Omega_i^t\|^2 \\
& - \frac{1}{4\gamma} R^t \left(1 - \frac{32L^2\beta^2}{\hat{\beta}^2\eta^2} \gamma^2 - \frac{96L^2}{\hat{\beta}^2\eta^2} \gamma^2 - \frac{24L^2}{\beta^2} \gamma^2 \right) + b^2 \left(\frac{2\beta^2\gamma}{\hat{\beta}\eta} + \frac{8\gamma\beta^3}{\hat{\beta}^2\eta^2} \right) + \hat{c}^2 \frac{2\gamma\beta}{G} \\
& + \frac{4\gamma\beta}{G\hat{\beta}\eta} (1 - \hat{\beta}\eta)^2 \sum_{i \in \mathcal{G}} \langle (g_i^t - v_i^t) + \beta(v_i^t - \nabla f_i(x^t)) + \beta(\nabla f_i(x^t) - \nabla f_i(x^{t+1})), \theta_i^{t+1} \rangle \\
& + \frac{16\gamma\beta^2}{G\hat{\beta}^2\eta^2} (1 - \beta) \sum_{i \in \mathcal{G}} \langle v_i^t - \nabla f_i(x^t), \theta_i^{t+1} \rangle + 4\gamma(1 - \beta) \langle v^t - \nabla f(x^t), \theta^{t+1} \rangle \\
& + \frac{16\gamma\beta^2}{G\hat{\beta}^2\eta^2} (1 - \beta) \sum_{i \in \mathcal{G}} \langle \nabla f_i(x^t) - \nabla f_i(x^{t+1}), \theta_i^{t+1} \rangle \\
& + 4\gamma(1 - \beta) \langle \nabla f(x^t) - \nabla f(x^{t+1}), \theta^{t+1} \rangle.
\end{aligned}$$

Using step-size restriction (c), and assumption 6 of the induction, we get rid of the term with R^t and obtain

$$\begin{aligned}
\Phi^{t+1} \leq & \Phi^t - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 + b^2 \left(\frac{2\beta^2\gamma}{\hat{\beta}\eta} + \frac{8\gamma\beta^3}{\hat{\beta}^2\eta^2} \right) + \hat{c}^2 \frac{2\gamma\beta}{G} + 2\gamma\hat{\beta}^2 a^2 + 16\gamma\hat{\beta}^2 c\delta_{\text{byz}} \tau^2 T^2 + 16\gamma\hat{\beta}^2 c\delta_{\text{byz}} \hat{a}^2 \\
& + \frac{4\gamma\beta}{G\hat{\beta}\eta} (1 - \hat{\beta}\eta)^2 \sum_{i \in \mathcal{G}} \langle (g_i^t - v_i^t) + \beta(v_i^t - \nabla f_i(x^t)) + \beta(\nabla f_i(x^t) - \nabla f_i(x^{t+1})), \theta_i^{t+1} \rangle \\
& + \frac{16\gamma\beta^2}{G\hat{\beta}^2\eta^2} (1 - \beta) \sum_{i \in \mathcal{G}} \langle v_i^t - \nabla f_i(x^t), \theta_i^{t+1} \rangle + 4\gamma(1 - \beta) \langle v^t - \nabla f(x^t), \theta^{t+1} \rangle \\
& + \frac{16\gamma\beta^2}{G\hat{\beta}^2\eta^2} (1 - \beta) \sum_{i \in \mathcal{G}} \langle \nabla f_i(x^t) - \nabla f_i(x^{t+1}), \theta_i^{t+1} \rangle \\
& + 4\gamma(1 - \beta) \langle \nabla f(x^t) - \nabla f(x^{t+1}), \theta^{t+1} \rangle.
\end{aligned}$$

Now we sum all the inequalities above for $t \in \{0, \dots, K\}$ and get

$$\begin{aligned}
\Phi^{K+1} \leq & \Phi^0 - \frac{\gamma}{2} \sum_{t=0}^K \|\nabla f(x^t)\|^2 + Kb^2 \left(\frac{2\beta^2\gamma}{\hat{\beta}\eta} + \frac{8\gamma\beta^3}{\hat{\beta}^2\eta^2} \right) + K\hat{c}^2 \frac{2\gamma\beta}{G} + 2\gamma\hat{\beta}^2 a^2 K + 16\gamma\hat{\beta}^2 c\delta_{\text{byz}} \hat{a}^2 K \\
& + 16\gamma\hat{\beta}^2 c\delta_{\text{byz}} \tau^2 T^2 K \\
& + \frac{4\gamma\beta}{G\hat{\beta}\eta} (1 - \hat{\beta}\eta)^2 \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle (g_i^t - v_i^t) + \beta(v_i^t - \nabla f_i(x^t)) + \beta(\nabla f_i(x^t) - \nabla f_i(x^{t+1})), \theta_i^{t+1} \rangle \\
& + \frac{16\gamma\beta^2}{G\hat{\beta}^2\eta^2} (1 - \beta) \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle v_i^t - \nabla f_i(x^t), \theta_i^{t+1} \rangle + 4\gamma(1 - \beta) \sum_{t=0}^K \langle v^t - \nabla f(x^t), \theta^{t+1} \rangle \\
& + \frac{16\gamma\beta^2}{G\eta^2} (1 - \beta) \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \nabla f_i(x^t) - \nabla f_i(x^{t+1}), \theta_i^{t+1} \rangle \\
& + 4\gamma(1 - \beta) \sum_{t=0}^K \langle \nabla f(x^t) - \nabla f(x^{t+1}), \theta^{t+1} \rangle. \tag{70}
\end{aligned}$$

Rearranging terms and using step-size restriction (b), we get

$$\begin{aligned}
\frac{\gamma}{2} \sum_{t=0}^K \|\nabla f(x^t)\|^2 &\leq \Phi^0 - \Phi^{K+1} + Kb^2 \left(\frac{2\beta^2\gamma}{\hat{\beta}\eta} + \frac{8\gamma\beta^3}{\hat{\beta}^2\eta^2} \right) + K\hat{c}^2 \frac{2\gamma\beta}{G} \\
&+ \frac{\beta}{6L} \hat{\beta}^2 a^2 K + \frac{4\beta}{3L} \hat{\beta}^2 c \delta_{\text{byz}} \hat{a}^2 K + \frac{4\beta}{3L} \hat{\beta}^2 \tau^2 c \delta_{\text{byz}} T^2 K \\
&+ \frac{4\gamma\beta}{G\hat{\beta}\eta} (1 - \hat{\beta}\eta)^2 \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle (g_i^t - v_i^t) + \beta(v_i^t - \nabla f_i(x^t)) + \beta(\nabla f_i(x^t) - \nabla f_i(x^{t+1})), \theta_i^{t+1} \rangle \\
&+ \frac{16\gamma\beta^2}{G\hat{\beta}^2\eta^2} (1 - \beta) \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle v_i^t - \nabla f_i(x^t), \theta_i^{t+1} \rangle + 4\gamma(1 - \beta) \sum_{t=0}^K \langle v^t - \nabla f(x^t), \theta^{t+1} \rangle \\
&+ \frac{16\gamma\beta^2}{G\hat{\beta}^2\eta^2} (1 - \beta) \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \nabla f_i(x^t) - \nabla f_i(x^{t+1}), \theta_i^{t+1} \rangle \\
&+ 4\gamma(1 - \beta) \sum_{t=0}^K \langle \nabla f(x^t) - \nabla f(x^{t+1}), \theta^{t+1} \rangle.
\end{aligned}$$

Taking into account that $\frac{\gamma}{2} \sum_{t=0}^K \|\nabla f(x^t)\|^2 \geq 0$ and using momentum restriction (a), we get that the event $E^K \cap \left(\cap_{i \in \mathcal{G}} \overline{\Theta}_i^{K+1} \right) \cap \overline{N}^{K+1} \cap \overline{\Theta}^{K+1} \cap \left(\cap_{i \in \mathcal{G}} \overline{N}_i^{K+1} \right)$ implies

$$\begin{aligned}
\Phi^{K+1} &\leq \Phi^0 + Kb^2 \left(\frac{2\beta^2\gamma}{\hat{\beta}\eta} + \frac{8\gamma\beta^3}{\hat{\beta}^2\eta^2} \right) + K\hat{c}^2 \frac{2\gamma\beta}{G} + \frac{\tau\hat{\beta}^3 a^2 K}{324L\sqrt{L}\Delta} + \frac{2\tau\hat{\beta}^3 c \delta_{\text{byz}} \hat{a}^2 K}{81L\sqrt{L}\Delta} + \frac{2\tau^3 \hat{\beta}^3 c \delta_{\text{byz}} T^2 K}{81L\sqrt{L}\Delta} \\
&+ \frac{4\gamma\beta}{G\hat{\beta}\eta} (1 - \hat{\beta}\eta)^2 \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle (g_i^t - v_i^t) + \beta(v_i^t - \nabla f_i(x^t)) + \beta(\nabla f_i(x^t) - \nabla f_i(x^{t+1})), \theta_i^{t+1} \rangle \\
&+ \frac{16\gamma\beta^2}{G\hat{\beta}^2\eta^2} (1 - \beta) \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle v_i^t - \nabla f_i(x^t), \theta_i^{t+1} \rangle + \frac{4\gamma(1 - \beta)}{G} \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle v^t - \nabla f(x^t), \theta_i^{t+1} \rangle \\
&+ \frac{16\gamma\beta^2}{G\hat{\beta}^2\eta^2} (1 - \beta) \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \nabla f_i(x^t) - \nabla f_i(x^{t+1}), \theta_i^{t+1} \rangle \\
&+ \frac{4\gamma(1 - \beta)}{G} \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \nabla f(x^t) - \nabla f(x^{t+1}), \theta_i^{t+1} \rangle.
\end{aligned}$$

Next, we define the following random vectors:

$$\begin{aligned}
\zeta_{1,i}^t &:= \begin{cases} g_i^t - v_i^t, & \text{if } \|g_i^t - v_i^t\| \leq B_{\text{init}}, \\ 0, & \text{otherwise} \end{cases}, \\
\zeta_{2,i}^t &:= \begin{cases} v_i^t - \nabla f_i(x^t), & \text{if } \|v_i^t - \nabla f_i(x^t)\| \leq \sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T + \hat{\beta}a \\ & + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a}, \\ 0, & \text{otherwise} \end{cases}, \\
\zeta_{3,i}^t &:= \begin{cases} \nabla f_i(x^t) - \nabla f_i(x^{t+1}), & \text{if } \|\nabla f_i(x^t) - \nabla f_i(x^{t+1})\| \leq L\gamma \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\hat{\beta}a \right. \\ & \left. + 3\sqrt{2c\delta_{\text{byz}}}\hat{a}\hat{\beta} + 5\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right), \\ 0, & \text{otherwise} \end{cases}, \\
\zeta_4^t &:= \begin{cases} v^t - \nabla f(x^t), & \text{if } \|v^t - \nabla f(x^t)\| \leq \sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T + \hat{\beta}a \\ & + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a}, \\ 0, & \text{otherwise} \end{cases}, \\
\zeta_5^t &:= \begin{cases} \nabla f(x^t) - \nabla f(x^{t+1}), & \text{if } \|\nabla f(x^t) - \nabla f(x^{t+1})\| \leq L\gamma \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\hat{\beta}a \right. \\ & \left. + 3\sqrt{2c\delta_{\text{byz}}}\hat{a}\hat{\beta} + 5\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right), \\ 0, & \text{otherwise} \end{cases}.
\end{aligned}$$

By definition, all introduced random vectors $\zeta_{l,i}^t, l \in [3], i \in \mathcal{G}, \zeta_{4,5}^t$ are bounded with probability 1. Moreover, by the definition of E^t we get that the event $E^K \cap \bar{\Theta}^{K+1} \cap \left(\cap_{i \in \mathcal{G}} \bar{\Theta}_i^{K+1} \right) \cap \bar{N}^{K+1} \cap \left(\cap_{i \in \mathcal{G}} \bar{N}_i^{K+1} \right)$ implies

$$\begin{aligned}
\zeta_{1,i}^t &= g_i^t - v_i^t, \quad \zeta_{2,i}^t = v_i^t - \nabla f_i(x^t), \quad \zeta_{3,i}^t = \nabla f_i(x^t) - \nabla f_i(x^{t+1}), \\
\zeta_4^t &= v^t - \nabla f(x^t), \quad \zeta_5^t = \nabla f(x^t) - \nabla f(x^{t+1}).
\end{aligned}$$

Therefore, the event $E^K \cap \bar{\Theta}^{K+1} \cap \left(\cap_{i \in \mathcal{G}} \bar{\Theta}_i^{K+1} \right) \cap \bar{N}^{K+1} \cap \left(\cap_{i \in \mathcal{G}} \bar{N}_i^{K+1} \right)$ implies

$$\begin{aligned}
\Phi^{K+1} &\leq \underbrace{Kb^2 \left(\frac{2\beta^2\gamma}{\hat{\beta}\eta} + \frac{8\gamma\beta^3}{\hat{\beta}^2\eta^2} \right)}_{\textcircled{1}} + \underbrace{\frac{4\gamma\beta}{G\hat{\beta}\eta}(1-\eta)^2 \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_{1,i}^t, \theta_i^{t+1} \rangle}_{\textcircled{2}} \\
&+ \underbrace{\frac{4\gamma\beta^2}{G\hat{\beta}\eta}(1-\hat{\beta}\eta)^2 \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_{2,i}^t, \theta_i^{t+1} \rangle}_{\textcircled{3}} + \underbrace{\frac{4\gamma\beta^2}{G\hat{\beta}\eta}(1-\hat{\beta}\eta)^2 \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_{3,i}^t, \theta_i^{t+1} \rangle}_{\textcircled{4}} \\
&+ \underbrace{\frac{16\gamma\beta^2}{G\hat{\beta}^2\eta^2}(1-\beta) \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_{2,i}^t, \theta_i^{t+1} \rangle}_{\textcircled{5}} + \underbrace{\frac{4\gamma(1-\beta)}{G} \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_4^t, \theta_i^{t+1} \rangle}_{\textcircled{6}} \\
&+ \underbrace{\frac{16\gamma\beta^2}{G\hat{\beta}^2\eta^2}(1-\beta) \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_{3,i}^t, \theta_i^{t+1} \rangle}_{\textcircled{7}} + \underbrace{\frac{4\gamma(1-\beta)}{G} \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_5^t, \theta_i^{t+1} \rangle}_{\textcircled{8}} \\
&+ \underbrace{\frac{\tau\hat{\beta}^3 a^2 K}{324L\sqrt{L\Delta}}}_{\textcircled{9}} + \underbrace{\frac{2\hat{\beta}^3 c\delta_{\text{byz}}\hat{a}^2 K}{81L\sqrt{L\Delta}} + \frac{2\tau^3\hat{\beta}^3 c\delta_{\text{byz}}T^2 K}{81L\sqrt{L\Delta}}}_{\textcircled{10}}.
\end{aligned}$$

Bound of the term ①. Since $12L\gamma \leq \beta$, for the term ① we have

$$Kb^2 \left(\frac{2\beta^2\gamma}{\hat{\beta}\eta} + \frac{8\gamma\beta^3}{\hat{\beta}^2\eta^2} \right) + K\hat{c}^2 \frac{2\gamma\beta}{G} \leq Kb^2 \left(\frac{\beta^3}{6L\hat{\beta}\eta} + \frac{2\beta^4}{3L\hat{\beta}^2\eta^2} \right) + K\hat{c}^2 \frac{\beta^2}{6LG}.$$

By choosing β such that

$$\beta \leq \min \left\{ \left(\frac{L\Delta\hat{\beta}\eta}{12Tb^2} \right)^{1/3}, \left(\frac{L\Delta\hat{\beta}^2\eta^2}{48Tb^2} \right)^{1/4}, \left(\frac{L\Delta G}{12T\hat{c}^2} \right)^{1/2} \right\} \quad (71)$$

we get that

$$Kb^2 \left(\frac{2\beta^2\gamma}{\hat{\beta}\eta} + \frac{8\gamma\beta^3}{\hat{\beta}^2\eta^2} \right) + K\hat{c}^2 \frac{2\gamma\beta}{G} \leq 3 \cdot \frac{\Delta}{72} = \frac{\Delta}{24}.$$

This bound holds with probability 1. Note that the worst dependency in the restriction on β w.r.t. T is $\mathcal{O}(1/T^{3/4})$ since $\hat{\beta} \sim \frac{1}{a} \sim \frac{1}{T}$ that comes from the second term in (71).

Bound of the term ②. For term ②, let us enumerate random variables as

$$\langle \zeta_{1,1}^0, \theta_1^1 \rangle, \dots, \langle \zeta_{1,G}^0, \theta_G^1 \rangle, \langle \zeta_{1,1}^1, \theta_1^2 \rangle, \dots, \langle \zeta_{1,G}^1, \theta_G^2 \rangle, \dots, \langle \zeta_{1,1}^K, \theta_1^{K+1} \rangle, \dots, \langle \zeta_{1,G}^K, \theta_G^{K+1} \rangle,$$

i.e., first by index i , then by index t . Then we have that the event $E^K \cap \left(\cap_{i \in \mathcal{G}} \bar{\Theta}_i^{K+1} \right)$ implies

$$\mathbb{E} \left[\frac{4\gamma\beta}{G\hat{\beta}\eta} (1-\eta)^2 \langle \zeta_{1,i}^l, \theta_i^{l+1} \rangle \mid \langle \zeta_{1,i-1}^l, \theta_{i-1}^{l+1} \rangle, \dots, \langle \zeta_{1,1}^l, \theta_1^{l+1} \rangle, \dots, \langle \zeta_{1,1}^0, \theta_1^1 \rangle \right] = 0,$$

because $\{\theta_i^{l+1}\}_{i \in \mathcal{G}}$ are independent. Let

$$\sigma_2^2 := \frac{16\gamma^2\beta^2}{G^2\hat{\beta}^2\eta^2} \cdot B_{\text{init}}^2 \cdot \sigma^2.$$

Since θ_i^{l+1} is σ -sub-Gaussian random vector, for

$$\mathbb{E} [\cdot \mid l, i-1] := \mathbb{E} [\cdot \mid \langle \zeta_{1,i-1}^l, \theta_{i-1}^{l+1} \rangle, \dots, \langle \zeta_{1,1}^l, \theta_1^{l+1} \rangle, \dots, \langle \zeta_{1,1}^0, \theta_1^1 \rangle]$$

we have

$$\begin{aligned} & \mathbb{E} \left[\exp \left(\left| \frac{1}{\sigma_2^2} \frac{16\gamma^2\beta^2}{G^2\hat{\beta}^2\eta^2} (1-\eta)^4 \langle \zeta_{1,i}^l, \theta_i^{l+1} \rangle^2 \right| \right) \mid l, i-1 \right] \\ & \leq \mathbb{E} \left[\exp \left(\frac{1}{\sigma_1^2} \frac{16\gamma^2\beta^2}{G^2\hat{\beta}^2\eta^2} \|\zeta_{1,i}^l\|^2 \cdot \|\theta_i^{l+1}\|^2 \right) \mid l, i-1 \right] \\ & \leq \mathbb{E} \left[\exp \left(\frac{1}{\sigma_2^2} \frac{16\gamma^2\beta^2}{G^2\hat{\beta}^2\eta^2} \cdot B_{\text{init}}^2 \|\theta_i^{l+1}\|^2 \right) \mid l, i-1 \right] \\ & \leq \mathbb{E} \left[\exp \left(\frac{G^2\hat{\beta}^2\eta^2}{16\gamma^2\beta^2 \cdot B_{\text{init}}^2 \cdot \sigma^2} \frac{16\gamma^2\beta^2}{G^2\hat{\beta}^2\eta^2} \cdot B_{\text{init}}^2 \|\theta_i^{l+1}\|^2 \right) \mid l, i-1 \right] \\ & = \mathbb{E} \left[\exp \left(\frac{\|\theta_i^{l+1}\|^2}{\sigma^2} \mid l, i-1 \right) \right] \leq \exp(1). \end{aligned}$$

Therefore, we have by Lemma B.2 with $\sigma_k^2 \equiv \sigma_2^2$ that

$$\begin{aligned} & \mathbb{P} \left(\frac{4\gamma\beta}{G\hat{\beta}\eta} (1-\hat{\beta}\eta)^2 \left\| \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_{1,i}^t, \theta_i^{t+1} \rangle \right\| \geq (\sqrt{2} + \sqrt{2}b_1) \sqrt{\sum_{t=0}^K \sum_{i \in \mathcal{G}} \frac{16B_{\text{init}}^2\gamma^2\beta^2\sigma^2}{G^2\hat{\beta}^2\eta^2}} \right) \\ & \leq \exp(-b_1^2/3) \\ & = \frac{\alpha}{20(T+1)} \end{aligned}$$

with $b_1^2 = 3 \log \left(\frac{20(T+1)}{\alpha} \right)$. Note that since $12L\gamma \leq \beta$

$$\begin{aligned} (\sqrt{2} + \sqrt{2}b_1) \sqrt{\sum_{t=0}^K \sum_{i \in \mathcal{G}} \frac{16B_{\text{init}}^2 \gamma^2 \beta^2 \sigma^2}{G^2 \hat{\beta}^2 \eta^2}} &\leq (\sqrt{2} + \sqrt{2}b_1) \sqrt{\sum_{t=0}^K \sum_{i \in \mathcal{G}} \frac{B_{\text{init}}^2 \beta^4 \sigma^2}{9L^2 G^2 \hat{\beta}^2 \eta^2}} \\ &= (\sqrt{2} + \sqrt{2}b_1) \frac{B_{\text{init}} \beta^2 \sigma}{3LG \hat{\beta} \eta} \sqrt{(K+1)G} \\ &\leq \frac{\Delta}{8}, \end{aligned}$$

because we choose β such that

$$\beta \leq \left(\frac{3L\Delta\sqrt{G}\hat{\beta}\eta}{8\sqrt{2}(1+b_1)B_{\text{init}}\sigma\sqrt{T}} \right)^{1/2}, \quad \text{and} \quad K+1 \leq T. \quad (72)$$

This implies that

$$\mathbb{P} \left(\frac{4\gamma\beta}{G\hat{\beta}\eta} (1 - \hat{\beta}\eta)^2 \left\| \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_{1,i}^t, \theta_i^{t+1} \rangle \right\| \geq \frac{\Delta}{8} \right) \leq \frac{\alpha}{20(T+1)}$$

with this choice of momentum parameter. The dependency of (72) on T is $\tilde{\mathcal{O}}(1/T^{3/4})$ since $\hat{\beta} \sim \frac{1}{T}$.

Bound of the term ③. The bound in this case is similar to the previous one. Let

$$\sigma_3^2 := \frac{16\gamma^2 \beta^4}{G^2 \hat{\beta}^2 \eta^2} \cdot \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T + \frac{3}{2}b + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} \right)^2 \cdot \sigma^2.$$

Then,

$$\begin{aligned} &\mathbb{E} \left[\exp \left(\left| \frac{1}{\sigma_3^2} \frac{16\gamma^2 \beta^4}{G^2 \hat{\beta}^2 \eta^2} (1 - \hat{\beta}\eta)^4 \langle \zeta_{2,i}^l, \theta_i^{l+1} \rangle^2 \right| \right) \mid l, i-1 \right] \\ &\leq \mathbb{E} \left[\exp \left(\frac{1}{\sigma_3^2} \frac{16\gamma^2 \beta^4}{G^2 \hat{\beta}^2 \eta^2} \|\zeta_{2,i}^l\|^2 \cdot \|\theta_i^{l+1}\|^2 \right) \right] \\ &\leq \mathbb{E} \left[\exp \left(\frac{1}{\sigma_3^2} \frac{16\gamma^2 \beta^4}{G^2 \hat{\beta}^2 \eta^2} \cdot \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T + \hat{\beta}a \right. \right. \right. \\ &\quad \left. \left. \left. + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} \right)^2 \cdot \|\theta_i^{l+1}\|^2 \right) \mid l, i-1 \right] \\ &\leq \mathbb{E} \left[\exp \left(\left[\frac{16\gamma^2 \beta^4}{G^2 \hat{\beta}^2 \eta^2} \cdot \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} \right)^2 \cdot \sigma^2 \right]^{-1} \times \right. \right. \\ &\quad \left. \left. \frac{16\gamma^2 \beta^4}{G^2 \hat{\beta}^2 \eta^2} \cdot \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} \right)^2 \cdot \|\theta_i^{l+1}\|^2 \right) \mid l, i-1 \right] \\ &= \mathbb{E} \left[\exp \left(\frac{\|\theta_i^{l+1}\|^2}{\sigma^2} \right) \mid l, i-1 \right] \leq \exp(1). \end{aligned}$$

Therefore, we have by Lemma B.2 that

$$\begin{aligned} &\mathbb{P} \left[\frac{4\gamma\beta^2}{G\hat{\beta}\eta} (1 - \hat{\beta}\eta)^2 \left\| \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_{2,i}^t, \theta_i^{t+1} \rangle \right\| \right. \\ &\geq (\sqrt{2} + \sqrt{2}b_1) \sqrt{\sum_{t=0}^K \sum_{i \in \mathcal{G}} \frac{16\gamma^2 \beta^4 \sigma^2}{G^2 \hat{\beta}^2 \eta^2} \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} \right)^2} \\ &\leq \exp(-b_1^2/3) = \frac{\alpha}{20(T+1)}. \end{aligned}$$

Note that by using the restrictions $\hat{\beta} \leq \min \left\{ \frac{\sqrt{L\Delta}}{a}, \frac{\sqrt{L\Delta}}{4\sqrt{2c\delta_{\text{byz}}\hat{a}}} \right\}$ and $12L\gamma \leq \beta$ we get

$$\begin{aligned}
& (\sqrt{2} + \sqrt{2}b_1)\sqrt{(K+1)G}\frac{4\gamma\beta^2\sigma}{\hat{\beta}\eta G} \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}\hat{a}} \right) \\
& \leq (\sqrt{2} + \sqrt{2}b_1)\sqrt{(K+1)G}\frac{\beta^3\sigma}{3L\hat{\beta}\eta G} \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T + 2\sqrt{L\Delta} \right) \\
& = (\sqrt{2} + \sqrt{2}b_1)\sqrt{(K+1)G}\frac{\beta^3\sigma}{3L\hat{\beta}\eta G} \left(4\sqrt{L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b \right) + \\
& \quad (1+b_1)\sqrt{(K+1)G}\frac{\beta^3\sigma}{L\eta G}\sqrt{c\delta_{\text{byz}}}\tau T \\
& \leq \frac{\Delta}{16} + \frac{\Delta}{16} = \frac{\Delta}{8}
\end{aligned}$$

holds because we choose

$$\beta \leq \min \left\{ \left(\frac{3L\Delta\hat{\beta}\eta\sqrt{G}}{16\sqrt{2}(1+b_1)\sigma\sqrt{T} \left(4\sqrt{L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b \right)} \right)^{1/3}, \left(\frac{L\Delta\eta\sqrt{G}}{16(1+b_1)\sqrt{c\delta_{\text{byz}}}\sigma\tau T^{3/2}} \right)^{1/3} \right\}, \quad (73)$$

and $K+1 \leq T$.

This implies

$$\mathbb{P} \left(\frac{4\gamma\beta^2}{G\hat{\beta}\eta} (1 - \hat{\beta}\eta)^2 \left\| \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_{2,i}^t, \theta_i^{t+1} \rangle \right\| \geq \frac{\Delta}{8} \right) \leq \frac{\alpha}{20(T+1)}.$$

Note that both terms in the choice of β have the dependency w.r.t. T of order $\tilde{\mathcal{O}}(1/T^{1/2})$ since $\hat{\beta} \sim \frac{1}{T}$.

Bound of the term ④. The bound in this case is similar to the previous one. Let

$$\sigma_4^2 := \frac{16L^2\gamma^4\beta^4}{G^2\hat{\beta}^2\eta^2} \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\hat{\beta}a + 3\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + 5\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2 \cdot \sigma^2.$$

Then we have

$$\begin{aligned}
& \mathbb{E} \left[\exp \left(\left| \frac{1}{\sigma_4^2} \frac{16\gamma^2\beta^4}{G^2\hat{\beta}^2\eta^2} (1 - \hat{\beta}\eta)^4 \langle \zeta_{3,i}^l, \theta_i^{l+1} \rangle^2 \right| \mid l, i - 1 \right) \right] \\
& \leq \mathbb{E} \left[\exp \left(\frac{1}{\sigma_4^2} \frac{16\gamma^2\beta^4}{G^2\hat{\beta}^2\eta^2} \|\zeta_{3,i}^l\|^2 \cdot \|\theta_i^{l+1}\|^2 \mid l, i - 1 \right) \right] \\
& \leq \mathbb{E} \left[\exp \left(\frac{1}{\sigma_4^2} \frac{16\gamma^2\beta^4}{G^2\hat{\beta}^2\eta^2} \cdot L^2\gamma^2 \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\hat{\beta}a + 3\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} \right. \right. \right. \\
& \quad \left. \left. \left. + 5\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2 \cdot \|\theta_i^{l+1}\|^2 \right) \mid l, i - 1 \right] \\
& \leq \mathbb{E} \left[\exp \left(\left[\frac{16L^2\gamma^4\beta^4}{G^2\hat{\beta}^2\eta^2} \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\hat{\beta}a + 3\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + 5\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2 \cdot \sigma^2 \right]^{-1} \times \right. \right. \\
& \quad \left. \left. \frac{16L^2\gamma^4\beta^4}{G^2\hat{\beta}^2\eta^2} \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\hat{\beta}a + 3\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + 5\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2 \cdot \|\theta_i^{l+1}\|^2 \right) \mid l, i - 1 \right] \\
& = \mathbb{E} \left[\exp \left(\frac{\|\theta_i^{l+1}\|^2}{\sigma^2} \right) \right] \leq \exp(1).
\end{aligned}$$

Therefore, we have by Lemma B.2 that

$$\begin{aligned}
& \mathbb{P} \left(\frac{4\gamma\beta^2}{G\hat{\beta}\eta} (1 - \hat{\beta}\eta)^2 \left\| \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_{3,i}^t, \theta_i^{t+1} \rangle \right\| \right. \\
& \geq \sqrt{2}(1 + b_1) \sqrt{\sum_{t=0}^K \sum_{i \in \mathcal{G}} \frac{16L^2\gamma^4\beta^4\sigma^2}{G^2\hat{\beta}^2\eta^2} \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau + b) + 3\hat{\beta}a + 3\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + 5\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2} \\
& \leq \exp(-b_1^2/3) = \frac{\alpha}{20(T+1)}.
\end{aligned}$$

Using the restrictions $\hat{\beta} \leq \min \left\{ \frac{\sqrt{L\Delta}}{a}, \frac{\sqrt{L\Delta}}{4\sqrt{2c\delta_{\text{byz}}}\hat{a}} \right\}$ and $12L\gamma \leq \beta$ we get

$$\begin{aligned}
& \sqrt{2}(1 + b_1) \sqrt{(K+1)G} \frac{4L\gamma^2\beta^2\sigma}{\hat{\beta}\eta G} \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau + b) + 3\hat{\beta}a + 3\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + 5\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right) \\
& \leq \sqrt{2}(1 + b_1) \sqrt{(K+1)G} \frac{\beta^4\sigma}{36L\hat{\beta}\eta G} \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau + b) + 4\sqrt{L\Delta} + 5\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right) \\
& = \sqrt{2}(1 + b_1) \sqrt{(K+1)G} \frac{\beta^4\sigma}{36L\hat{\beta}\eta G} \left(12\sqrt{L\Delta} + 3(B_{\text{init}} - \tau + b) \right) + \\
& \quad (1 + b_1) \sqrt{(K+1)G} \frac{5\beta^4\sigma}{18L\eta G} \sqrt{c\delta_{\text{byz}}}\tau T \\
& \leq \frac{\Delta}{16} + \frac{\Delta}{16} = \frac{\Delta}{8},
\end{aligned}$$

because we choose β such that

$$\beta \leq \min \left\{ \left(\frac{9L\Delta\hat{\beta}\eta\sqrt{G}}{16\sqrt{2}(1 + b_1)\sigma\sqrt{T} \left(12\sqrt{L\Delta} + 3(B_{\text{init}} - \tau + b) \right)} \right)^{1/4}, \left(\frac{9L\Delta\eta\sqrt{G}}{80(1 + b_1)\sigma\sqrt{c\delta_{\text{byz}}}T^{3/2}\tau} \right)^{1/4} \right\}, \quad (74)$$

and $K + 1 \leq T$.

This implies

$$\mathbb{P} \left(\frac{4\gamma\beta^2}{G\hat{\beta}\eta} (1 - \hat{\beta}\eta)^2 \left\| \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_{2,i}^t, \theta_i^{t+1} \rangle \right\| \geq \frac{\Delta}{8} \right) \leq \frac{\alpha}{20(T+1)},$$

Note that both terms in the choice of β have the dependency w.r.t. T of order $\tilde{\mathcal{O}}(1/T^{3/8})$ since $\hat{\beta} \sim \frac{1}{T}$.

Bound of the term ⑤. The bound in this case is similar to the previous one. Let

$$\sigma_5^2 := \frac{256\gamma^2\beta^4}{G^2\hat{\beta}^4\eta^4} \cdot \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2 \cdot \sigma^2.$$

Then we have

$$\begin{aligned}
& \mathbb{E} \left[\exp \left(\left| \frac{1}{\sigma_5^2} \frac{256\gamma^2\beta^4}{G^2\hat{\beta}^4\eta^4} (1-\beta)^2 \langle \zeta_{2,i}^l, \theta_i^{l+1} \rangle^2 \right| \mid l, i-1 \right) \right] \\
& \leq \mathbb{E} \left[\exp \left(\frac{1}{\sigma_5^2} \frac{256\gamma^2\beta^4}{G^2\hat{\beta}^4\eta^4} \|\zeta_{2,i}^l\|^2 \cdot \|\theta_i^{l+1}\|^2 \right) \mid l, i-1 \right] \\
& \leq \mathbb{E} \left[\exp \left(\frac{1}{\sigma_5^2} \frac{256\gamma^2\beta^4}{G^2\hat{\beta}^4\eta^4} \cdot \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau + b) + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \right. \right. \right. \\
& \quad \left. \left. \left. \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2 \cdot \|\theta_i^{l+1}\|^2 \right) \mid l, i-1 \right] \\
& = \mathbb{E} \left[\exp \left(\left[\frac{256\gamma^2\beta^4}{L^2G^2\hat{\beta}^4\eta^4} \cdot \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau + b) + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2 \cdot \sigma^2 \right]^{-1} \right. \right. \\
& \quad \left. \left. \frac{256\gamma^2\beta^4}{G^2\hat{\beta}^4\eta^4} \cdot \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2 \cdot \|\theta_i^{l+1}\|^2 \right) \mid l, i-1 \right] \\
& = \mathbb{E} \left[\exp \left(\frac{\|\theta_i^{l+1}\|^2}{\sigma^2} \right) \mid l, i-1 \right] \leq \exp(1).
\end{aligned}$$

Therefore, we have by Lemma B.2 that

$$\begin{aligned}
& \mathbb{P} \left[\frac{16\gamma\beta^2}{G\hat{\beta}^2\eta^2} (1-\beta) \left\| \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_{2,i}^t, \theta_i^{t+1} \rangle \right\| \right] \\
& \geq (\sqrt{2} + \sqrt{2}b_1) \sqrt{\sum_{t=0}^K \sum_{i \in \mathcal{G}} \frac{256\gamma^2\beta^4\sigma^2}{G^2\hat{\beta}^4\eta^4} \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau + b) + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2} \\
& \leq \exp(-b_1^2/3) = \frac{\alpha}{20(T+1)}.
\end{aligned}$$

Using the restrictions $12L\gamma \leq \beta$ and $\hat{\beta} \leq \left\{ \frac{\sqrt{L\Delta}}{a}, \frac{\sqrt{L\Delta}}{4\sqrt{2c\delta_{\text{byz}}}\hat{a}} \right\}$ we get

$$\begin{aligned}
& \sqrt{2}(1+b_1)\sqrt{(K+1)G} \frac{16\gamma\beta^2\sigma}{G\hat{\beta}^2\eta^2} \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right) \\
& \leq \sqrt{2}(1+b_1)\sqrt{(K+1)G} \frac{4\beta^3\sigma}{3LG\hat{\beta}^2\eta^2} \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \sqrt{L\Delta} + \sqrt{L\Delta} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right) \\
& = \sqrt{2}(1+b_1)\sqrt{(K+1)G} \frac{4\beta^3\sigma}{3LG\hat{\beta}^2\eta^2} \left(4\sqrt{L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau + b) \right) + \\
& \quad (1+b_1)\sqrt{(K+1)G} \frac{4\beta^3\sigma}{3LG\hat{\beta}^2\eta^2} \sqrt{c\delta_{\text{byz}}}\tau T \\
& \leq \frac{\Delta}{16} + \frac{\Delta}{16} = \frac{\Delta}{8}
\end{aligned}$$

because we choose β such that

$$\beta \leq \min \left\{ \left(\frac{3L\Delta\hat{\beta}^2\eta^2\sqrt{G}}{64\sqrt{2}(1+b_1)\sigma\sqrt{T} \left(4\sqrt{L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau + b) \right)} \right)^{1/3}, \left(\frac{3L\Delta\hat{\beta}^2\eta^2\sqrt{G}}{32(1+b_1)\tau T^{3/2}\sqrt{c\delta_{\text{byz}}}} \right)^{1/3} \right\}, \quad (75)$$

and $K+1 \leq T$.

This implies

$$\mathbb{P} \left(\frac{16\gamma\beta^2}{G\hat{\beta}^2\eta^2} (1-\hat{\beta}\beta) \left\| \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_{2,i}^t, \theta_i^{t+1} \rangle \right\| \geq \frac{\Delta}{8} \right) \leq \frac{\alpha}{20(T+1)}.$$

Note that both terms in the choice of β have the dependency w.r.t. T of order $\tilde{\mathcal{O}}(1/T^{5/6})$ since $\hat{\beta} \sim \frac{1}{T}$.

Bound of the term ⑦. The bound in this case is similar to the previous one. Let

$$\sigma_7^2 := \frac{256L^2\gamma^4\beta^4}{G^2\hat{\beta}^4\eta^4} \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau + b) + 3\hat{\beta}a + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\hat{a} + 5\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2 \cdot \sigma^2.$$

Then we have

$$\begin{aligned} & \mathbb{E} \left[\exp \left(\left| \frac{1}{\sigma_7^2} \frac{256L^2\gamma^4\beta^4}{G^2\hat{\beta}^4\eta^4} (1-\beta)^2 \langle \zeta_{3,i}^l, \theta_i^{l+1} \rangle^2 \right| \mid l, i-1 \right) \right] \\ & \leq \mathbb{E} \left[\exp \left(\frac{1}{\sigma_7^2} \frac{256\gamma^2\beta^4}{G^2\hat{\beta}^4\eta^4} \|\zeta_{3,i}^l\|^2 \cdot \|\theta_i^{l+1}\|^2 \right) \mid l, i-1 \right] \\ & \leq \mathbb{E} \left[\exp \left(\frac{256\gamma^2\beta^4}{G^2\hat{\beta}^4\eta^4} \cdot L^2\gamma^2 \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau + b) + 3\hat{\beta}a + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\hat{a} + \right. \right. \right. \\ & \quad \left. \left. \left. 4\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2 \cdot \|\theta_i^{l+1}\|^2 \right) \mid l, i-1 \right] \\ & \leq \mathbb{E} \left[\exp \left(\left[\frac{256L^2\gamma^4\beta^4}{G^2\hat{\beta}^4\eta^4} \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau + b) + 3\hat{\beta}a + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\hat{a} + 4\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2 \cdot \sigma^2 \right]^{-1} \times \right. \right. \\ & \quad \left. \left. \frac{256L^2\gamma^4\beta^4}{G^2\hat{\beta}^4\eta^4} \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau + b) + 3\hat{\beta}a + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\hat{a} + 4\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2 \cdot \|\theta_i^{l+1}\|^2 \right) \mid l, i-1 \right] \\ & = \mathbb{E} \left[\exp \left(\frac{\|\theta_i^{l+1}\|^2}{\sigma^2} \right) \mid l, i-1 \right] \leq \exp(1). \end{aligned}$$

Therefore, we have by Lemma B.2 that

$$\begin{aligned} & \mathbb{P} \left[\frac{16\gamma\beta^2}{G\hat{\beta}^2\eta^2} (1-\beta) \left\| \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_{3,i}^t, \theta_i^{t+1} \rangle \right\| \geq \right. \\ & \quad \left. \sqrt{2}(1+b_1) \sqrt{\sum_{t=0}^K \sum_{i \in \mathcal{G}} \frac{256L^2\gamma^4\beta^4\sigma^2}{G^2\hat{\beta}^4\eta^4} \cdot \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau + b) + 3\hat{\beta}a + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\hat{a} + 4\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2} \right] \\ & \leq \exp(-b_1^2/3) = \frac{\alpha}{20(T+1)}. \end{aligned}$$

Using the restrictions $12L\gamma \leq \beta$ and $\hat{\beta} \leq \min \left\{ \frac{\sqrt{L\Delta}}{a}, \frac{\sqrt{L\Delta}}{4\sqrt{2c\delta_{\text{byz}}}\hat{a}} \right\}$ we get

$$\begin{aligned} & \sqrt{2}(1+b_1) \sqrt{(K+1)G} \frac{16L\gamma^2\beta^2\sigma}{\hat{\beta}^2\eta^2G} \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau + b) + 3\hat{\beta}a + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\hat{a} + 4\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right) \\ & \leq \sqrt{2}(1+b_1) \sqrt{(K+1)G} \frac{\beta^4\sigma}{9L\hat{\beta}^2\eta^2G} \left(8\sqrt{L\Delta} + 3(B_{\text{init}} - \tau + b) + 3\sqrt{L\Delta} + \frac{3}{4}\sqrt{L\Delta} + 4\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right) + \\ & \quad \sqrt{2}(1+b_1) \sqrt{(K+1)G} \frac{\beta^4\sigma}{9L\hat{\beta}^2\eta^2G} \left(4\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right) \\ & = \sqrt{2}(1+b_1) \sqrt{(K+1)G} \frac{\beta^4\sigma}{9L\hat{\beta}^2\eta^2G} \left(12\sqrt{L\Delta} + 3(B_{\text{init}} - \tau + b) \right) + \\ & \quad + \sqrt{2}(1+b_1) \sqrt{(K+1)G} \frac{4\beta^4\sigma}{9L\hat{\beta}\eta^2G} \sqrt{2c\delta_{\text{byz}}}\tau T \\ & \leq \frac{\Delta}{16} + \frac{\Delta}{16} = \frac{\Delta}{8} \end{aligned}$$

because we choose

$$\beta \leq \min \left\{ \left(\frac{9L\Delta\hat{\beta}^2\eta^2\sqrt{G}}{16\sqrt{2}(1+b_1)\sigma\sqrt{T}\left(12\sqrt{L\Delta}+3(B_{\text{init}}-\tau+b)\right)} \right)^{1/4}, \left(\frac{9L\Delta\hat{\beta}^2\eta^2\sqrt{G}}{128(1+b_1)\sqrt{c\delta_{\text{byz}}}\tau\sigma T^{3/2}} \right)^{1/4} \right\}, \quad (76)$$

and $K+1 \leq T$.

This implies

$$\mathbb{P} \left(\frac{8\gamma\beta^2}{G\eta^2}(1-\beta) \left\| \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_{3,i}^t, \theta_i^{t+1} \rangle \right\| \geq \frac{\Delta}{8} \right) \leq \frac{\alpha}{20(T+1)}.$$

Note that both terms in the choice of β have the dependency w.r.t. T of order $\tilde{\mathcal{O}}(1/T^{5/8})$ since $\hat{\beta} \sim \frac{1}{T}$.

Bound of the term ⑥. The bound in this case is similar to the previous one. Let

$$\sigma_6^2 := \frac{16\gamma^2}{G^2} \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau + b) + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2 \cdot \sigma^2.$$

Then we have

$$\begin{aligned} & \mathbb{E} \left[\exp \left(\left| \frac{1}{\sigma_6^2} \frac{16\gamma^2}{G^2} (1-\beta)^2 \langle \zeta_4^l, \theta_i^{l+1} \rangle^2 \right| \right) \mid l, i-1 \right] \\ & \leq \mathbb{E} \left[\exp \left(\frac{1}{\sigma_6^2} \frac{16\gamma^2}{G^2} \|\zeta_4^l\|^2 \cdot \|\theta_i^{l+1}\|^2 \right) \mid l, i-1 \right] \\ & \leq \mathbb{E} \left[\exp \left(\frac{1}{\sigma_6^2} \frac{16\gamma^2}{G^2} \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau + b) + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \right. \right. \right. \\ & \quad \left. \left. \left. \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2 \cdot \|\theta_i^{l+1}\|^2 \right) \mid l, i-1 \right] \\ & \leq \mathbb{E} \left[\exp \left(\left[\frac{16\gamma^2}{G^2} \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau + b) + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2 \cdot \sigma^2 \right]^{-1} \right. \right. \\ & \quad \left. \left. \frac{16\gamma^2}{G^2} \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau) + \frac{3}{2}b + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2 \cdot \|\theta_i^{l+1}\|^2 \right) \mid l, i-1 \right] \\ & = \mathbb{E} \left[\exp \left(\frac{\|\theta_i^{l+1}\|^2}{\sigma^2} \right) \mid l, i-1 \right] \leq \exp(1). \end{aligned}$$

Therefore, we have by Lemma B.2 that

$$\begin{aligned} & \mathbb{P} \left[\frac{4\gamma(1-\beta)}{G} \left\| \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_{4,i}^t, \theta_i^{t+1} \rangle \right\| \right. \\ & \geq \sqrt{2}(1+b_1) \sqrt{\sum_{t=0}^K \sum_{i \in \mathcal{G}} \frac{16\gamma^2}{G^2} \sigma^2 \cdot \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau + b) + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}}\hat{a} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2} \\ & \leq \exp(-b_1^2/3) = \frac{\alpha}{20(T+1)}, \end{aligned}$$

Using the restrictions $12L\gamma \leq \beta$ and $\hat{\beta} \leq \min \left\{ \frac{\sqrt{L\Delta}}{a}, \frac{\sqrt{L\Delta}}{4\sqrt{2c\delta_{\text{byz}}\hat{a}}} \right\}$ we get

$$\begin{aligned}
& (\sqrt{2} + \sqrt{2}b_1)\sqrt{(K+1)G} \cdot \frac{4\gamma}{G} \sigma \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau + b) + \hat{\beta}a + 4\hat{\beta}\sqrt{2c\delta_{\text{byz}}\hat{a}} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right) \\
& \leq (\sqrt{2} + \sqrt{2}b_1)\sqrt{(K+1)G} \cdot \frac{\beta}{3LG} \sigma \left(\sqrt{4L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau + b) + \sqrt{L\Delta} + \sqrt{L\Delta} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right) \\
& = (\sqrt{2} + \sqrt{2}b_1)\sqrt{(K+1)G} \cdot \frac{\beta}{3LG} \sigma \left(4\sqrt{L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau + b) \right) + \\
& \quad (\sqrt{2} + \sqrt{2}b_1)\sqrt{(K+1)G} \cdot \frac{\beta}{LG} \sigma \sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \\
& \leq \sqrt{2}(1+b_1)\sqrt{(K+1)} \cdot \frac{\beta}{3L\sqrt{G}} \sigma \left(4\sqrt{L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau + b) \right) + \\
& \quad (1+b_1)\sqrt{(K+1)} \cdot \frac{4}{29\sqrt{c\delta_{\text{byz}}TL\sqrt{G}}} \sigma \sqrt{c\delta_{\text{byz}}}\hat{\beta}\tau T \\
& \leq \sqrt{2}(1+b_1)\sqrt{(K+1)} \cdot \frac{\beta}{3L\sqrt{G}} \sigma \left(4\sqrt{L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau + b) \right) + (1+b_1)\sqrt{(K+1)} \frac{4}{29L\sqrt{G}} \sigma \hat{\beta}\tau \\
& \leq \frac{\Delta}{16} + \frac{\Delta}{16} = \frac{\Delta}{8},
\end{aligned}$$

where we use the restriction $\beta \leq \frac{2}{29\sqrt{c\delta_{\text{byz}}T}}$ and choose β and $\hat{\beta}$ such that

$$\begin{aligned}
\beta & \leq \left(\frac{3L\Delta\sqrt{G}}{16\sqrt{2}(1+b_1)\sigma\sqrt{T} \left(4\sqrt{L\Delta} + \frac{3}{2}(B_{\text{init}} - \tau + b) \right)} \right), \quad \hat{\beta} \leq \left(\frac{29L\Delta\sqrt{G}}{64(1+b_1)\sigma\sqrt{T}} \right), \\
& \text{and } K+1 \leq T.
\end{aligned} \tag{77}$$

This implies

$$\mathbb{P} \left(\frac{4\gamma(1-\beta)}{G} \left\| \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_{4,i}^t, \theta_i^{t+1} \rangle \right\| \geq \frac{\Delta}{8} \right) \leq \frac{\alpha}{20(T+1)}.$$

Note that the worst dependency in the choice of β and $\hat{\beta}$ w.r.t. T is $\tilde{\mathcal{O}}(1/T^{1/2})$.

Bound of the term ⑧. The bound in this case is similar to the previous one. Let

$$\sigma_8^2 := \frac{16L^2\gamma^4}{G^2} \cdot \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau + b) + 3\hat{\beta}a + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\hat{a} + 5\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2 \cdot \sigma^2.$$

Then we have

$$\begin{aligned}
& \mathbb{E} \left[\exp \left(\left| \frac{1}{\sigma_8^2} \frac{16\gamma^2}{G^2} (1-\beta)^2 \langle \zeta_5^l, \theta_i^{l+1} \rangle^2 \right| \mid l, i-1 \right) \right] \\
& \leq \mathbb{E} \left[\exp \left(\frac{1}{\sigma_8^2} \frac{16\gamma^2}{G^2} \|\zeta_5^l\|^2 \cdot \|\theta_i^{l+1}\|^2 \right) \mid l, i-1 \right] \\
& \leq \mathbb{E} \left[\exp \left(\frac{1}{\sigma_8^2} \frac{16\gamma^2}{G^2} L^2\gamma^2 \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau + b) + 3\hat{\beta}a + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\hat{a} + \right. \right. \right. \\
& \quad \left. \left. \left. 5\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right) \cdot \|\theta_i^{l+1}\|^2 \right) \mid l, i-1 \right].
\end{aligned}$$

Since θ_i^{l+1} is sub-Gaussian with parameter σ^2 , then we can continue the chain of inequalities above using the definition of σ_8^2

$$\begin{aligned} & \mathbb{E} \left[\exp \left(\left[\frac{16L^2\gamma^4}{G^2} \cdot \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau + b) + 3\hat{\beta}a + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\hat{a} + 5\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2 \cdot \sigma^2 \right]^{-1} \right. \right. \\ & \quad \left. \left. \frac{16L^2\gamma^4}{G^2} \cdot \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau + b) + 3\hat{\beta}a + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\hat{a} + 5\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2 \cdot \|\theta_i^{l+1}\|^2 \right) \mid l, i-1 \right] \\ & = \mathbb{E} \left[\exp \left(\frac{\|\theta_i^{l+1}\|^2}{\sigma^2} \right) \right] \leq \exp(1). \end{aligned}$$

Therefore, we have by Lemma B.2 that

$$\begin{aligned} & \Pr \left[\frac{4\gamma(1-\beta)}{G} \left\| \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_{5,i}^t, \theta^{t+1} \rangle \right\| \right. \\ & \geq (\sqrt{2} + \sqrt{2}b_1) \sqrt{\sum_{t=0}^K \sum_{i \in \mathcal{G}} \frac{16L^2\gamma^4}{G^2} \sigma^2 \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau + b) + 3\hat{\beta}a + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\hat{a} + 5\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right)^2} \\ & \left. \leq \exp(-b_1^2/3) = \frac{\alpha}{20(T+1)} \right]. \end{aligned}$$

Using the restrictions $12L\gamma \leq \beta$ and $\hat{\beta} \leq \min \left\{ \frac{\sqrt{L\Delta}}{a}, \frac{\sqrt{L\Delta}}{4\sqrt{2c\delta_{\text{byz}}}\hat{a}} \right\}$ we get

$$\begin{aligned} & (\sqrt{2} + \sqrt{2}b_1) \sqrt{(K+1)G} \cdot \frac{4L\gamma^2}{G} \sigma \left(\sqrt{64L\Delta} + 3(B_{\text{init}} - \tau + b) + 3\hat{\beta}a + 3\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\hat{a} + 5\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right) \\ & \leq (\sqrt{2} + \sqrt{2}b_1) \sqrt{(K+1)G} \cdot \frac{\beta^2\sigma}{36LG} \left(8\sqrt{L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\sqrt{L\Delta} + \frac{3}{4}\sqrt{L\Delta} + 5\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right) \\ & \leq \sqrt{2}(1+b_1) \sqrt{(K+1)G} \cdot \frac{\beta^2\sigma}{36LG} \left(12\sqrt{L\Delta} + 3(B_{\text{init}} - \tau + b) \right) + \\ & \quad (1+b_1) \sqrt{(K+1)G} \cdot \frac{5\beta^2\sigma}{18LG} \sqrt{c\delta_{\text{byz}}}\hat{\beta}\tau T \\ & \leq (\sqrt{2} + \sqrt{2}b_1) \sqrt{(K+1)G} \cdot \frac{\beta^2\sigma}{36LG} \left(8\sqrt{L\Delta} + 3(B_{\text{init}} - \tau) + 3b + 3\sqrt{L\Delta} + \frac{3}{4}\sqrt{L\Delta} + 5\sqrt{2c\delta_{\text{byz}}}\hat{\beta}\tau T \right) \\ & \leq \sqrt{2}(1+b_1) \sqrt{(K+1)G} \cdot \frac{\beta^2\sigma}{36LG} \left(12\sqrt{L\Delta} + 3(B_{\text{init}} - \tau + b) \right) + \\ & \quad (1+b_1) \sqrt{(K+1)G} \cdot \frac{64 \cdot 5\tau^2\sigma\hat{\beta}^2}{431^2 \cdot 18L^2\Delta G} \sqrt{c\delta_{\text{byz}}}\hat{\beta}\tau T \\ & \leq \frac{\Delta}{16} + \frac{\Delta}{16} = \frac{\Delta}{8} \end{aligned}$$

because we choose β and $\hat{\beta}$ such that

$$\beta \leq \left(\frac{9L\Delta\sqrt{G}}{4\sqrt{2}(1+b_1)\sigma\sqrt{T} \left(12\sqrt{L\Delta} + 3(B_{\text{init}} - \tau + b) \right)} \right)^{1/2}, \quad \hat{\beta} \leq \left(\frac{462 \cdot (L\Delta)^2\sqrt{G}}{(1+b_1)\sigma T^{3/2}\tau^3} \right)^{1/2} \quad (78)$$

and $K+1 \leq T$.

This implies

$$\mathbb{P} \left(4\gamma(1-\beta) \left\| \sum_{t=0}^K \sum_{i \in \mathcal{G}} \langle \zeta_{5,i}^t, \theta^{t+1} \rangle \right\| \geq \frac{\Delta}{8} \right) \leq \frac{\alpha}{20(T+1)}.$$

Note that the worst dependency in the choice of β and $\hat{\beta}$ w.r.t T is $\tilde{O}(1/T^{3/4})$.

Bound of the term ⑨. We need to choose hyperparameters such that

$$\frac{\tau \hat{\beta}^3 a^2 K}{324 L \sqrt{L \Delta}} \leq \frac{\Delta}{24}.$$

This is possible by choosing $\hat{\beta}$ such that

$$\hat{\beta} \leq \left(\frac{27(L\Delta)^{3/2}}{2a^2 T \tau} \right)^{1/3} \quad \text{and} \quad K \leq T + 1. \quad (79)$$

Note that since $a \sim T$ we obtain $\hat{\beta} \sim \frac{1}{T}$.

Bound of the term ⑩. We need to choose hyperparameters such that

$$\frac{2\tau \hat{\beta}^3 c \delta_{\text{byz}} \hat{a}^2 K}{81 L \sqrt{L \Delta}} + \frac{2\tau^3 \hat{\beta}^3 c \delta_{\text{byz}} T^2 K}{81 L \sqrt{L \Delta}} \leq \frac{\Delta}{24}.$$

By choosing $\hat{\beta}$ such that

$$\hat{\beta} \leq \min \left\{ \left(\frac{81(L\Delta)^{3/2}}{96 c \delta_{\text{byz}} \hat{a}^2 \tau T} \right)^{1/3}, \left(\frac{81(L\Delta)^{3/2}}{96 c \delta_{\text{byz}} \tau^3 T^3} \right)^{1/3} \right\}, \quad \text{and} \quad K \leq T + 1 \quad (80)$$

we have

$$\frac{2\tau \hat{\beta}^3 c \delta_{\text{byz}} \hat{a}^2 K}{81 L \sqrt{L \Delta}} + \frac{2\tau^3 \hat{\beta}^3 c \delta_{\text{byz}} T^2 K}{81 L \sqrt{L \Delta}} \leq \frac{\Delta}{48} + \frac{\Delta}{48} = \frac{\Delta}{24}.$$

Final probability. Therefore, the probability event

$$\Omega := E^K \cap \bar{\Theta}^{K+1} \cap \left(\cap_{i \in \mathcal{G}} \bar{\Theta}_i^{K+1} \right) \cap \bar{N}^{K+1} \cap \left(\cap_{i \in \mathcal{G}} \bar{N}_i^{K+1} \right) \cap E_{\textcircled{1}} \cap E_{\textcircled{2}} \cap E_{\textcircled{3}} \cap E_{\textcircled{4}} \cap E_{\textcircled{5}} \cap E_{\textcircled{6}} \cap E_{\textcircled{7}} \cap E_{\textcircled{8}} \cap E_{\textcircled{9}} \cap E_{\textcircled{10}},$$

where each $E_{\textcircled{1}}-E_{\textcircled{10}}$ denotes that each of 2-8-th terms is smaller than $\frac{\Delta}{8}$ and each of the terms 1, 9, 10 is smaller than $\frac{\Delta}{24}$. This implies that

$$\textcircled{1} + \textcircled{2} + \textcircled{3} + \textcircled{4} + \textcircled{5} + \textcircled{6} + \textcircled{7} + \textcircled{8} + \textcircled{9} + \textcircled{10} \leq 7 \cdot \frac{\Delta}{8} + 3 \cdot \frac{\Delta}{24} = \Delta,$$

i.e., condition 7 in the induction assumption holds. Moreover, this also implies that

$$\Phi^{K+1} \leq \Phi^0 + \Delta \leq \Delta + \Delta = 2\Delta,$$

i.e., condition 6 in the induction assumption holds. The probability $\mathbb{P}(E_{K+1})$ can be lower bounded as follows

$$\begin{aligned} \mathbb{P}(E_{K+1}) &\geq \mathbb{P}(\Omega) \\ &= \mathbb{P} \left(E^K \cap \bar{\Theta}^{K+1} \cap \left(\cap_{i \in \mathcal{G}} \bar{\Theta}_i^{K+1} \right) \cap \bar{N}^{K+1} \cap \left(\cap_{i \in \mathcal{G}} \bar{N}_i^{K+1} \right) \cap E_{\textcircled{1}} \cap E_{\textcircled{2}} \cap E_{\textcircled{3}} \cap E_{\textcircled{4}} \cap E_{\textcircled{5}} \cap E_{\textcircled{6}} \right. \\ &\quad \left. \cap E_{\textcircled{7}} \cap E_{\textcircled{8}} \cap E_{\textcircled{9}} \cap E_{\textcircled{10}} \right) \\ &= 1 - \mathbb{P} \left(\bar{E}_K \cup \bar{\Theta}^{K+1} \cup \left(\cup_{i \in \mathcal{G}} \bar{\Theta}_i^{K+1} \right) \cup \bar{N}^{K+1} \cup \left(\cup_{i \in \mathcal{G}} \bar{N}_i^{K+1} \right) \cup \bar{E}_{\textcircled{1}} \cup \bar{E}_{\textcircled{2}} \cup \bar{E}_{\textcircled{3}} \cup \bar{E}_{\textcircled{4}} \cup \bar{E}_{\textcircled{5}} \cup \bar{E}_{\textcircled{6}} \right. \\ &\quad \left. \cup \bar{E}_{\textcircled{7}} \cup \bar{E}_{\textcircled{8}} \cup \bar{E}_{\textcircled{9}} \cup \bar{E}_{\textcircled{10}} \right) \\ &\geq 1 - \mathbb{P}(\bar{E}_K) - \mathbb{P}(\bar{\Theta}^{K+1}) - \sum_{i \in \mathcal{G}} \mathbb{P}(\bar{\Theta}_i^{K+1}) - \mathbb{P}(\bar{N}^{K+1}) - \sum_{i \in \mathcal{G}} \mathbb{P}(\bar{N}_i^{K+1}) - \mathbb{P}(\bar{E}_{\textcircled{1}}) - \mathbb{P}(\bar{E}_{\textcircled{2}}) \\ &\quad - \mathbb{P}(\bar{E}_{\textcircled{3}}) - \mathbb{P}(\bar{E}_{\textcircled{4}}) - \mathbb{P}(\bar{E}_{\textcircled{5}}) - \mathbb{P}(\bar{E}_{\textcircled{6}}) - \mathbb{P}(\bar{E}_{\textcircled{7}}) - \mathbb{P}(\bar{E}_{\textcircled{8}}) - \mathbb{P}(\bar{E}_{\textcircled{9}}) - \mathbb{P}(\bar{E}_{\textcircled{10}}) \\ &\geq 1 - \frac{\alpha(K+1)}{T+1} - \frac{\alpha}{8(T+1)} - \sum_{i \in \mathcal{G}} \frac{\alpha}{8G(T+1)} - \frac{\alpha}{8(T+1)} - 0 - 10 \cdot \frac{\alpha}{20(T+1)} \\ &= 1 - \frac{\alpha(K+2)}{T+1}. \end{aligned}$$

This finalizes the transition step of induction. The result of the theorem follows by setting $K = T - 1$. Indeed, from (70) we obtain

$$\frac{\gamma}{2} \sum_{t=0}^K \|\nabla f(x^t)\|^2 \leq \Phi^0 - \Phi^{K+1} + \Delta \leq 2\Delta \Rightarrow \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|^2 \leq \frac{4\Delta}{\gamma T}. \quad (81)$$

Final rate. We highlight that we are interested in the functional dependency of the rate on the problem constants. Therefore, in the rest of the proof we omit using numerical constants. Translating momentum restrictions (71), (72), (73), (74), (75), (77), (76), and (78) to the step-size restriction using $12L\gamma = \beta$ equality we get that the step-size should satisfy

$$\gamma = \frac{1}{L} \tilde{\mathcal{O}} \left(\min \left\{ \underbrace{\left(\frac{L\Delta G}{T\sigma^2} \right)^{1/2}}_{\text{from term 1 (71)}}, \underbrace{\left(\frac{L\Delta \hat{\beta}^2 \eta^2}{T\sigma^2} \right)^{1/4}}_{\text{from term 2 (72)}}, \underbrace{\left(\frac{L\Delta \hat{\beta} \eta}{T\sigma^2} \right)^{1/3}}_{\text{from term 2 (72)}}, \underbrace{\left(\frac{L\Delta \sqrt{G} \hat{\beta} \eta}{B_{\text{init}} \sigma \sqrt{T}} \right)^{1/2}}_{\text{from term 2 (72)}}, \right. \\ \underbrace{\left(\frac{L\Delta \sqrt{G} \hat{\beta} \eta}{\sigma(\sqrt{L\Delta} + B_{\text{init}} + \sigma)\sqrt{T}} \right)^{1/3}}_{\text{from term 3 (73)}}, \underbrace{\left(\frac{L\Delta \sqrt{G} \eta}{\sigma \sqrt{c\delta_{\text{byz}} \tau T^{3/2}}} \right)^{1/3}}_{\text{from term 3 (73)}}, \underbrace{\left(\frac{L\Delta \hat{\beta} \eta \sqrt{G}}{\sigma(\sqrt{L\Delta} + B_{\text{init}} + \sigma)\sqrt{T}} \right)^{1/4}}_{\text{from term 4 (74)}}, \underbrace{\left(\frac{L\Delta \eta \sqrt{G}}{\sigma \sqrt{c\delta_{\text{byz}} T^{3/2} \tau}} \right)^{1/4}}_{\text{from term 4 (74)}}, \\ \underbrace{\left(\frac{L\Delta \hat{\beta}^2 \eta^2 \sqrt{G}}{\sigma(\sqrt{L\Delta} + B_{\text{init}} + \sigma)\sqrt{T}} \right)^{1/3}}_{\text{from term 5 (75)}}, \underbrace{\left(\frac{L\Delta \hat{\beta} \eta^2 \sqrt{G}}{\sigma T^{3/2} \sqrt{c\delta_{\text{byz}} \tau}} \right)^{1/3}}_{\text{from term 5 (75)}}, \\ \underbrace{\left(\frac{L\Delta \hat{\beta}^2 \eta^2 \sqrt{G}}{\sigma(\sqrt{L\Delta} + B_{\text{init}} + \sigma)\sqrt{T}} \right)^{1/4}}_{\text{from term 7 (76)}}, \underbrace{\left(\frac{L\Delta \hat{\beta} \eta^2 \sqrt{G}}{\sqrt{c\delta_{\text{byz}} \sigma \tau T^{3/2}}} \right)^{1/4}}_{\text{from term 7 (76)}}, \\ \left. \underbrace{\left(\frac{L\Delta \sqrt{G}}{\sigma(\sqrt{L\Delta} + B_{\text{init}} + \sigma)\sqrt{T}} \right)}_{\text{from term 6 (77)}}, \underbrace{\left(\frac{L\Delta \sqrt{G}}{\sigma(\sqrt{L\Delta} + B_{\text{init}} + \sigma)\sqrt{T}} \right)^{1/2}}_{\text{from term 8 (78)}} \right\} \right), \quad (82)$$

and

$$\hat{\beta} \leq \tilde{\mathcal{O}} \left(\underbrace{\left(\frac{L\Delta \sqrt{G}}{\sigma \sqrt{T}} \right)}_{\text{from term 6 (77)}}, \underbrace{\left(\frac{(L\Delta)^2 \sqrt{G}}{\sigma T^{3/2} \tau^3} \right)^{1/2}}_{\text{from term 8 (78)}}, \underbrace{\frac{\sqrt{L\Delta}}{a^{2/3} T^{1/3} \tau^{1/3}}}_{\text{from term 9 (79)}}, \underbrace{\frac{\sqrt{L\Delta}}{(c\delta_{\text{byz}})^{1/3} \hat{a}^{2/3} \tau^{1/3} T^{1/3}}}_{\text{from term 10 (80)}}, \underbrace{\frac{\sqrt{L\Delta}}{(c\delta_{\text{byz}})^{1/3} \tau T}}_{\text{from term 10 (80)}} \right).$$

To guarantee the convergence to the neighborhood, we should choose $\hat{\beta} \sim \frac{1}{a} \sim \frac{1}{T}$. Therefore, the first two restrictions on $\hat{\beta}$ will not contribute to the size of the neighborhood. Moreover, the worst power of T in (82) comes from the term ⑤ and equals $\gamma \sim \frac{1}{T^{5/6}}$. The second worst comes from terms ① and ②, and equals to $\gamma \sim \frac{1}{T^{3/4}}$. These terms give the rate of the form

$$\frac{L\Delta}{T} \tilde{\mathcal{O}} \left(\left(\frac{T\sigma^2}{L\Delta \hat{\beta}^2 \eta^2} \right)^{1/4} + \left(\frac{B_{\text{init}} \sigma \sqrt{T}}{L\Delta \sqrt{G} \hat{\beta} \eta} \right)^{1/2} + \left(\frac{\sigma(\sqrt{L\Delta} + B_{\text{init}} + \sigma)\sqrt{T}}{L\Delta \sqrt{G} \hat{\beta}^2 \eta^2} \right)^{1/3} + \left(\frac{\sigma \sqrt{c\delta_{\text{byz}} \tau T^{3/2}}}{L\Delta \sqrt{G} \hat{\beta} \eta} \right)^{1/3} \right). \quad (83)$$

Now we need to plug in all restrictions on $\hat{\beta}$ in (83). We remind that in total we should satisfy (up to numerical constants)

$$\hat{\beta} \leq \sqrt{L\Delta} \cdot \tilde{\mathcal{O}} \left(\min \left\{ \frac{1}{a}, \frac{1}{\sqrt{c\delta_{\text{byz}} \hat{a}}}, \frac{1}{a^{2/3} T^{1/3} \tau^{1/3}}, \frac{1}{(c\delta_{\text{byz}})^{1/3} \hat{a}^{2/3} \tau^{1/3} T^{1/3}}, \frac{1}{(c\delta_{\text{byz}})^{1/3} \tau T} \right\} \right).$$

We can simplify this constraint to

$$\hat{\beta} \leq \sqrt{L\Delta} \cdot \tilde{\mathcal{O}} \left(\min \left\{ \frac{1}{aC}, \frac{1}{a^{2/3}T^{1/3}\tau^{1/3}}, \frac{1}{(c\delta_{\text{byz}})^{1/3}\hat{a}^{2/3}\tau^{1/3}T^{1/3}}, \frac{1}{(c\delta_{\text{byz}})^{1/3}\tau T} \right\} \right), \quad (84)$$

where $C := (1 + \sqrt{G}\sqrt{c\delta_{\text{byz}}})$. Let us do this one by one.

1. $\hat{\beta} \sim \frac{\sqrt{L\Delta}}{aC}$, where $a \sim \tilde{\mathcal{O}}(\sqrt{d}\sigma_\omega\sqrt{T}/\sqrt{G})$, gives the term in the rate of the form

$$\begin{aligned} & \frac{L\Delta}{T} \tilde{\mathcal{O}} \left(\left(\frac{T\sigma^2}{L\Delta\hat{\beta}^2\eta^2} \right)^{1/4} + \left(\frac{B_{\text{init}}\sigma\sqrt{T}}{L\Delta\sqrt{G}\hat{\beta}\eta} \right)^{1/2} + \left(\frac{\sigma(\sqrt{L\Delta} + B_{\text{init}} + \sigma)\sqrt{T}}{L\Delta\sqrt{G}\hat{\beta}^2\eta^2} \right)^{1/3} + \left(\frac{\sigma\sqrt{c\delta_{\text{byz}}}\tau T^{3/2}}{L\Delta\sqrt{G}\hat{\beta}\eta} \right)^{1/3} \right) \\ &= \frac{L\Delta}{T} \tilde{\mathcal{O}} \left(\left(\frac{B_{\text{init}}^2 C^2 a^2 T \sigma^2}{(L\Delta)^2 \tau^2} \right)^{1/4} + \left(\frac{B_{\text{init}}^2 C a \sigma T^{1/2}}{(L\Delta)^{3/2} \sqrt{G} \tau} \right)^{1/2} + \left(\frac{B_{\text{init}}^2 C^2 a^2 \sigma (\sqrt{L\Delta} + B_{\text{init}} + \sigma) \sqrt{T}}{(L\Delta)^2 \sqrt{G} \tau^2} \right)^{1/3} \right. \\ & \quad \left. + \left(\frac{B_{\text{init}} C a \sigma \sqrt{c\delta_{\text{byz}}} T^{3/2}}{(L\Delta)^{3/2} \sqrt{G}} \right)^{1/3} \right) \\ &= \tilde{\mathcal{O}} \left(\left(\frac{(L\Delta)^2 B_{\text{init}}^2 C^2 a^2 \sigma^2}{T^3 \tau^2} \right)^{1/4} + \left(\frac{\sqrt{L\Delta} B_{\text{init}}^2 C a \sigma}{T^{1/2} \sqrt{G} \tau} \right)^{1/2} + \left(\frac{L\Delta B_{\text{init}}^2 C^2 a^2 \sigma (\sqrt{L\Delta} + B_{\text{init}} + \sigma)}{T^{5/2} \sqrt{G} \tau^2} \right)^{1/3} \right. \\ & \quad \left. + \left(\frac{(L\Delta)^{3/2} B_{\text{init}} C a \sigma \sqrt{c\delta_{\text{byz}}}}{\sqrt{G} T^{3/2}} \right)^{1/3} \right). \end{aligned}$$

Now we use the specific form of $a \sim \sqrt{d}\sigma_\omega\sqrt{T}/\sqrt{G}$ to obtain the rate

$$\begin{aligned} & \tilde{\mathcal{O}} \left(\left(\frac{(L\Delta)^2 B_{\text{init}}^2 C^2 d \sigma_\omega^2 \sigma^2}{G T^2 \tau^2} \right)^{1/4} + \left(\frac{\sqrt{L\Delta} B_{\text{init}}^2 C \sqrt{d} \sigma_\omega \sigma}{G \tau} \right)^{1/2} \right. \\ & \quad \left. + \left(\frac{L\Delta B_{\text{init}}^2 C^2 d \sigma_\omega^2 \sigma (\sqrt{L\Delta} + B_{\text{init}} + \sigma)}{(TG)^{3/2} \tau^2} \right)^{1/3} + \left(\frac{(L\Delta)^{3/2} B_{\text{init}} C \sqrt{d} \sigma_\omega \sigma \sqrt{c\delta_{\text{byz}}} \tau}{TG \tau} \right)^{1/3} \right). \quad (85) \end{aligned}$$

Since $\sigma_\omega \sim \sqrt{T}$, we omit the first two terms in (85) as they have better dependency on T . The final rate in this regime is

$$\tilde{\mathcal{O}} \left(\left(\frac{L\Delta B_{\text{init}}^2 C^2 d \sigma_\omega^2 \sigma (\sqrt{L\Delta} + B_{\text{init}} + \sigma)}{(TG)^{3/2} \tau^2} \right)^{1/3} + \left(\frac{(L\Delta)^{3/2} B_{\text{init}} C \sqrt{d} \sigma_\omega \sigma \sqrt{c\delta_{\text{byz}}} \tau}{TG \tau} \right)^{1/3} \right). \quad (86)$$

2. $\hat{\beta} \sim \frac{\sqrt{L\Delta}}{C_1 a^{2/3} T^{1/3} \tau^{1/3}}$, where $C_1 = 1 + G^{1/3}(c\delta_{\text{byz}})^{1/3}$ (we combine cases 2 and 3 in (84) together to simplify calculations). We obtain

$$\begin{aligned} & \frac{L\Delta}{T} \tilde{\mathcal{O}} \left(\left(\frac{T\sigma^2}{L\Delta\hat{\beta}^2\eta^2} \right)^{1/4} + \left(\frac{B_{\text{init}}\sigma\sqrt{T}}{L\Delta\sqrt{G}\hat{\beta}\eta} \right)^{1/2} + \left(\frac{\sigma(\sqrt{L\Delta} + B_{\text{init}} + \sigma)\sqrt{T}}{L\Delta\sqrt{G}\hat{\beta}^2\eta^2} \right)^{1/3} + \left(\frac{\sigma\sqrt{c\delta_{\text{byz}}}\tau T^{3/2}}{L\Delta\sqrt{G}\hat{\beta}\eta} \right)^{1/3} \right) \\ &= \tilde{\mathcal{O}} \left(\left(\frac{(L\Delta)^2 B_{\text{init}}^2 C_1^2 a^{4/3} \sigma^2}{T^{7/3} \tau^{4/3}} \right)^{1/4} + \left(\frac{\sqrt{L\Delta} B_{\text{init}}^2 C_1 a^{2/3} \sigma}{T^{7/6} \sqrt{G} \tau^{2/3}} \right)^{1/2} \right. \\ & \quad \left. + \left(\frac{L\Delta B_{\text{init}}^2 C_1^2 a^{4/3} \sigma (\sqrt{L\Delta} + B_{\text{init}} + \sigma)}{T^{11/6} \sqrt{G} \tau^{4/3}} \right)^{1/3} + \left(\frac{(L\Delta)^{3/2} B_{\text{init}} C_1 a^{2/3} \sigma \sqrt{c\delta_{\text{byz}}} \tau}{T^{7/6} \sqrt{G} \tau^{2/3}} \right)^{1/3} \right). \end{aligned}$$

Now we use the specific form of $a \sim \sqrt{d}\sigma_\omega\sqrt{T}/\sqrt{G}$ to obtain the rate

$$\begin{aligned} \tilde{\mathcal{O}} & \left(\left(\frac{(L\Delta)^2 B_{\text{init}}^2 C_1^2 d^{2/3} \sigma_\omega^{4/3} \sigma^2}{G^{2/3} T^{5/3} \tau^{4/3}} \right)^{1/4} + \left(\frac{\sqrt{L\Delta} B_{\text{init}}^2 C_1 d^{1/3} \sigma_\omega^{2/3} \sigma}{(GT)^{5/6} \tau^{2/3}} \right)^{1/2} \right. \\ & \left. + \left(\frac{L\Delta B_{\text{init}}^2 C_1^2 d^{2/3} \sigma_\omega^{4/3} \sigma (\sqrt{L\Delta} + B_{\text{init}} + \sigma)}{(TG)^{7/6} \tau^{4/3}} \right)^{1/3} + \left(\frac{(L\Delta)^{3/2} B_{\text{init}} C_1 d^{1/3} \sigma_\omega^{2/3} \sigma \sqrt{c\delta_{\text{byz}} \tau}}{(GT)^{5/6} \tau^{2/3}} \right)^{1/3} \right). \end{aligned} \quad (87)$$

Since $\sigma_\omega \sim \sqrt{T}$, then the first term in (87) has better dependency on T than the other two, therefore we omit them. The final rate in this case is

$$\tilde{\mathcal{O}} \left(\left(\frac{L\Delta B_{\text{init}}^2 C_1^2 d^{2/3} \sigma_\omega^{4/3} \sigma (\sqrt{L\Delta} + B_{\text{init}} + \sigma)}{(TG)^{7/6} \tau^{4/3}} \right)^{1/3} + \left(\frac{(L\Delta)^{3/2} B_{\text{init}} C_1 d^{1/3} \sigma_\omega^{2/3} \sigma \sqrt{c\delta_{\text{byz}} \tau}}{(GT)^{5/6} \tau^{2/3}} \right)^{1/3} \right). \quad (88)$$

3. $\hat{\beta} \sim \frac{\sqrt{L\Delta}}{(c\delta_{\text{byz}})^{1/3} \tau T}$, gives the term in the rate of the form

$$\begin{aligned} & \frac{L\Delta}{T} \tilde{\mathcal{O}} \left(\left(\frac{T\sigma^2}{L\Delta \hat{\beta}^2 \eta^2} \right)^{1/4} + \left(\frac{B_{\text{init}} \sigma \sqrt{T}}{L\Delta \sqrt{G} \hat{\beta} \eta} \right)^{1/2} + \left(\frac{\sigma (\sqrt{L\Delta} + B_{\text{init}} + \sigma) \sqrt{T}}{L\Delta \sqrt{G} \hat{\beta}^2 \eta^2} \right)^{1/3} + \left(\frac{\sigma \sqrt{c\delta_{\text{byz}} \tau} T^{3/2}}{L\Delta \sqrt{G} \hat{\beta} \eta} \right)^{1/3} \right) \\ & = \tilde{\mathcal{O}} \left(\left(\frac{(L\Delta)^2 (c\delta_{\text{byz}})^{2/3} B_{\text{init}}^2 \sigma^2}{T} \right)^{1/4} + \left(\frac{(L\Delta)^{1/2} B_{\text{init}}^2 (c\delta_{\text{byz}})^{1/3} \sigma}{\sqrt{GT}} \right)^{1/2} \right. \\ & \quad \left. + \left(\frac{L\Delta B_{\text{init}}^2 (c\delta_{\text{byz}})^{2/3} \sigma (\sqrt{L\Delta} + B_{\text{init}} + \sigma)}{\sqrt{TG}} \right)^{1/3} + \left(\frac{(c\delta_{\text{byz}})^{5/6} (L\Delta)^{3/2} B_{\text{init}} \sigma \tau}{\sqrt{TG}} \right)^{1/3} \right). \end{aligned}$$

We keep only the last two terms, since they have the worst dependence on T . Therefore, the final rate in this case is

$$\tilde{\mathcal{O}} \left(\left(\frac{L\Delta B_{\text{init}}^2 \sigma (\sqrt{L\Delta} + B_{\text{init}} + \sigma)}{\sqrt{TG}} \right)^{1/3} + \left(\frac{\sqrt{c\delta_{\text{byz}}} (L\Delta)^{3/2} B_{\text{init}} \sigma \tau}{\sqrt{TG}} \right)^{1/3} \right). \quad (89)$$

We observe that, in all three cases above, we still have descent with T . Therefore, those terms do not contribute to the utility bound. Now we should consider the momentum constraints (a)-(d) that can be combined into one as

$$\beta \leq \tilde{\mathcal{O}} \left(\min \left\{ \frac{\hat{\beta} \tau}{\sqrt{L\Delta} + B_{\text{init}} + \sigma}, \frac{1}{\sqrt{c\delta_{\text{byz}} T}} \right\} \right).$$

This translates into the rate of the form

$$\tilde{\mathcal{O}} \left(\frac{L\Delta (\sqrt{L\Delta} + B_{\text{init}} + \sigma)}{T \hat{\beta} \tau} + L\Delta \sqrt{c\delta_{\text{byz}}} \right).$$

We should plug in the restrictions (84) in the rate above. This leads to the rate of the form

$$\begin{aligned} & \tilde{\mathcal{O}} \left(\frac{L\Delta a C (\sqrt{L\Delta} + B_{\text{init}} + \sigma)}{T \sqrt{L\Delta} \tau} + \frac{L\Delta C_1 a^{2/3} T^{1/3} \tau^{1/3} (\sqrt{L\Delta} + B_{\text{init}} + \sigma)}{T \sqrt{L\Delta} \tau} + \frac{L\Delta \tau T (\sqrt{L\Delta} + B_{\text{init}} + \sigma)}{T \sqrt{L\Delta} \tau} \right. \\ & \quad \left. + L\Delta \sqrt{c\delta_{\text{byz}}} \right) \\ & = \tilde{\mathcal{O}} \left(\frac{\sqrt{L\Delta} a C (\sqrt{L\Delta} + B_{\text{init}} + \sigma)}{T \tau} + \frac{\sqrt{L\Delta} C_1 a^{2/3} (\sqrt{L\Delta} + B_{\text{init}} + \sigma)}{T^{2/3} \tau^{2/3}} \right. \\ & \quad \left. + \sqrt{L\Delta} (\sqrt{L\Delta} + B_{\text{init}} + \sigma) ((c\delta_{\text{byz}})^{1/3} + \sqrt{c\delta_{\text{byz}}}) \right). \end{aligned}$$

Now we use $a \sim \frac{\sqrt{d}\sigma_\omega\sqrt{T}}{\sqrt{G}}$, $C = (1 + \sqrt{G})$, and $C_1 = (1 + G^{1/3}(c\delta_{\text{byz}})^{1/3})$ to obtain the rate

$$\begin{aligned}
& (\sqrt{L\Delta} + B_{\text{init}} + \sigma) \cdot \tilde{\mathcal{O}} \left(\frac{\sqrt{L\Delta}\sqrt{d}\sigma_\omega C}{\sqrt{GT}\tau} + \frac{\sqrt{L\Delta}C_1 d^{1/3}\sigma_\omega^{2/3}}{(TG)^{1/3}\tau^{2/3}} + \sqrt{L\Delta}((c\delta_{\text{byz}})^{1/3} + \sqrt{c\delta_{\text{byz}}}) \right) \\
&= (\sqrt{L\Delta} + B_{\text{init}} + \sigma) \cdot \tilde{\mathcal{O}} \left(\frac{\sqrt{L\Delta}\sqrt{d}\sigma_\omega(1 + \sqrt{c\delta_{\text{byz}}G})}{\sqrt{GT}\tau} + \frac{\sqrt{L\Delta}(1 + (c\delta_{\text{byz}}G)^{1/3})d^{1/3}\sigma_\omega^{2/3}}{(TG)^{1/3}\tau^{2/3}} \right. \\
&\quad \left. + \sqrt{L\Delta}((c\delta_{\text{byz}})^{1/3} + \sqrt{c\delta_{\text{byz}}}) \right) \\
&= (\sqrt{L\Delta} + B_{\text{init}} + \sigma) \tilde{\mathcal{O}} \left(\frac{\sqrt{L\Delta}d\sigma_\omega}{\sqrt{GT}\tau} + \frac{\sqrt{L\Delta}d^{1/3}\sigma_\omega^{2/3}}{(TG)^{1/3}\tau^{2/3}} + \frac{\sqrt{c\delta_{\text{byz}}L\Delta}d\sigma_\omega}{\sqrt{T}\tau} + \frac{\sqrt{L\Delta}(c\delta_{\text{byz}}d\sigma_\omega^2)^{1/3}}{T^{1/3}\tau^{2/3}} \right. \\
&\quad \left. + \sqrt{L\Delta}((c\delta_{\text{byz}})^{1/3} + \sqrt{c\delta_{\text{byz}}}) \right).
\end{aligned}$$

We can combine the terms above in a simpler way as follows

$$(\sqrt{L\Delta} + B_{\text{init}} + \sigma) \sqrt{L\Delta} \tilde{\mathcal{O}} \left(\left(\frac{\sqrt{d}\sigma_\omega}{\sqrt{GT}\tau} + \frac{\sqrt{c\delta_{\text{byz}}d}\sigma_\omega}{\sqrt{T}\tau} + \sqrt{c\delta_{\text{byz}}} \right) + \left(\frac{\sqrt{d}\sigma_\omega}{\sqrt{GT}\tau} + \frac{\sqrt{c\delta_{\text{byz}}d}\sigma_\omega}{\sqrt{T}\tau} + \sqrt{c\delta_{\text{byz}}} \right)^{2/3} \right).$$

We observe that the first term is purely due to DP noise, the middle term is a mix of DP noise and robust aggregation, and the last term is purely due to robust aggregation.

Finally, the restriction (68) translate to

$$\gamma \leq \frac{1}{L} \tilde{\mathcal{O}} \left(\frac{\hat{\beta}\tau}{LB_{\text{init}}}, \frac{\sqrt{\hat{\beta}\tau}}{L\sqrt{B_{\text{init}}}} \right). \quad (90)$$

Note that the second restriction in (90) does not contribute to the neighborhood. The first restriction in (90) is already taken into account by the restriction $\beta \leq \tilde{\mathcal{O}} \left(\frac{\hat{\beta}\tau}{\sqrt{L\Delta} + B_{\text{init}} + \sigma} \right)$.

Case $\mathcal{I}_{K+1} = 0$. This case is even easier. The only change will be with the term next to R^t . We will get

$$1 - \frac{96L^2}{\hat{\beta}^2\eta^2}\gamma^2 - \frac{24L^2}{\beta^2}\gamma^2 \geq \frac{1}{3} - \frac{96L^2}{\hat{\beta}^2\eta^2}\gamma^2 \geq 0$$

instead of

$$1 - \frac{32\beta^2L^2}{\hat{\beta}^2\eta^2}\gamma^2 - \frac{96L^2}{\hat{\beta}^2\eta^2}\gamma^2 - \frac{24L^2}{\beta^2}\gamma^2 \geq 0$$

as in the previous case. This difference comes from Lemma C.9 because $\tilde{V}^{K+1} = 0$. The rest is a repetition of the previous derivations.

Bounds on Δ , B_{init} , and Δ_0 . It remains to give explicit bounds on Δ and B_{init} using problem-dependent constants. With the initialization $v_i^0 = g_i^0 = 0$ for all $i \in \mathcal{G}$ we have that

$$\begin{aligned}
\Phi^0 &= f(x^0) - f^* + \frac{8\gamma\beta}{\hat{\beta}^2\eta^2} \frac{1}{G} \sum_{i \in \mathcal{G}} \|\nabla f_i(x^0)\|^2 + \frac{2\gamma}{\beta} \|\nabla f(x^0)\|^2 \\
&\stackrel{(i)}{\leq} f(x^0) - f^* + \frac{8\gamma\beta}{\hat{\beta}^2\eta^2} B_\zeta \|\nabla f(x^0)\|^2 + \frac{8\gamma\beta}{\hat{\beta}^2\eta^2} \zeta^2 + \frac{4\gamma}{\beta} L(f(x^0) - f^*) \\
&\stackrel{(ii)}{\leq} \left(1 + \frac{16L\gamma\beta}{\hat{\beta}^2\eta^2} B_\zeta + \frac{4L\gamma}{\beta} \right) (f(x^0) - f^*) + \frac{8\gamma\beta}{\hat{\beta}^2\eta^2} \zeta^2. \quad (91)
\end{aligned}$$

where (i) follow from Assumption 3.5a and L -smoothness, (ii) – from L -smoothness. Note that we choose $\beta = 12L\gamma$, which implies that $\frac{4L\gamma}{\beta} = \frac{1}{3}$, and from (68) we have

$$\frac{8\gamma\beta}{\hat{\beta}^2\eta^2} = \frac{96L\gamma^2}{\hat{\beta}^2\eta^2} \leq 96L \cdot \frac{5}{6 \cdot 96L^2} = \frac{5}{6L}.$$

Thus, we have the bound on Φ^0 from (91)

$$\Phi^0 \leq (4/3 + \frac{5}{3}B_\zeta)(f(x^0) - f^*) + \frac{5}{6L}\zeta^2.$$

Therefore, we have that

$$L\Delta = \mathcal{O}((1 + B_\zeta)L(f(x^0) - f^*) + \zeta^2) = \mathcal{O}((1 + B_\zeta)LF^0 + \zeta^2)$$

Next, we bound $B_{\text{init}} = \max\{3\tau, \max_{i \in \mathcal{G}} \|\nabla f_i(x^0)\| + b\} > \tau$. Note that in the theorem, we assume that $\max_{i \in \mathcal{G}} \|\nabla f_i(x^0)\| \geq 3\tau$. This implies that $B_{\text{init}} = \max_{i \in \mathcal{G}} \|\nabla f_i(x^0)\| + b$. Therefore, we provide the bound with high probability as follows

$$\begin{aligned} B_{\text{init}} &= b + \max_{i \in \mathcal{G}} \|\nabla f_i(x^0)\| \\ &= \tilde{\mathcal{O}}\left(\sigma + \max_{i \in \mathcal{G}} \|\nabla f_i(x^0) - \nabla f(x^0)\| + \|\nabla f(x^0)\|\right) \\ &= \tilde{\mathcal{O}}\left(\sigma + \sqrt{B_\zeta LF^0} + \zeta + \sqrt{LF^0}\right). \end{aligned}$$

where we again use Assumption 3.5a and L -smoothness. This allows to bound Δ_0 as follows

$$\begin{aligned} \Delta_0 &= \sqrt{L\Delta} \left(\sqrt{L\Delta} + B_{\text{init}} + \sigma \right) \\ &= \tilde{\mathcal{O}}\left((1 + \sqrt{B_\zeta})(\sqrt{LF^0} + \zeta)(\sigma + (1 + \sqrt{B_\zeta})\sqrt{LF^0} + \zeta)\right) \\ &= \tilde{\mathcal{O}}\left((1 + B_\zeta)LF^0 + \zeta^2 + \sigma((1 + \sqrt{B_\zeta})\sqrt{LF^0} + \zeta)\right). \end{aligned}$$

□

Theorem D.2 (Theorem 3.22 in [Dwork et al., 2014]). *Let $\varepsilon \in (0, 1)$ be arbitrary. For $c^2 > 2 \log(1.25/\delta)$, the Gaussian mechanism with parameter $\sigma_\omega \geq c\Delta_2(f)/\varepsilon$ is (ε, δ) -differentially private.*

Corollary D.3. *Under the setup of Theorem 5.1 (Theorem D.1), if we set $\sigma_\omega = \Theta\left(\frac{\tau}{\varepsilon} \sqrt{T \log(\frac{T}{\delta}) \log(\frac{1}{\delta})}\right)$ for some $\varepsilon, \delta \in (0, 1)$, then the full T -round transcript of each honest client in Byz-Clip21-SGD2M satisfies local (ε, δ) -DP.*

Proof. We need to plug in the choice of σ_ω to Theorem D.1 and omit all terms that decay with T . Note that each step of Byz-Clip21-SGD2M satisfies $(\tilde{\varepsilon}, \tilde{\delta})$ -local DP where

$$\tilde{\varepsilon} = \frac{\varepsilon}{2\sqrt{2T \log \frac{1}{\delta}}}, \quad \tilde{\delta} = \frac{\delta}{T},$$

by Theorem D.2. Combining the above result with advanced composition theorem ([Dwork et al., 2014, Theorem 3.20]), we obtain that all T iterations of Byz-Clip21-SGD2M satisfy local (ε, δ) -differentially private. □

E PROOFS IN SPECIAL CASES

E.1 CONVERGENCE IN THE ABSENCE OF BOTH DP AND BYZANTINE ADVERSARIES

Corollary E.1 (Full statement of Corollary 5.4). *Let Assumptions 3.3, 3.4, and 3.5a and define $B_{\text{init}} := \max\{3\tau, \max_i \{\|\nabla f_i(x^0)\|\} + b\}$. Let the failure probability α be such that $\alpha \in (0, 1)$, and constants b and $\hat{c}$ be defined as in (67), and $\Delta \geq \Phi^0$ for Φ^0 defined in (9). Consider the run of Byz-Clip21-SGD2M (Algorithm 1) for T iterations with no DP and Byzantine adversaries, i.e., $\sigma_\omega = 0$ and $\delta_{\text{byz}} = 0$. Assume the following inequalities hold*

1. **step-size restrictions:**

$$(a) \quad \gamma \leq \frac{1}{24L};$$

(b) $12L\gamma = \beta$;

(c)

$$\frac{5}{6} - \frac{32L^2\beta^2}{\hat{\beta}^2\eta^2}\gamma^2 - \frac{96L^2}{\hat{\beta}^2\eta^2}\gamma^2 \geq 0. \quad (92)$$

2. momentum restrictions:

(a) $\hat{\beta} = 1$;

(b) $\beta \leq \min \left\{ \frac{8\tau}{431\sqrt{L\Delta}}, 1 \right\}$;

(c) $\beta \leq \min \left\{ \frac{2\tau}{39(B_{\text{init}} - \tau)}, 1 \right\}$;

(d) $\beta \leq \min \left\{ \frac{2\tau}{69b}, 1 \right\}$;

(e) and momentum restrictions defined in (71), (72), (73), (74), (75), (77), (76), (78), (79), and (80) with $\hat{\beta} = 1$, $a, \hat{a} = 0, \delta_{\text{byz}} = 0$;

Then, with probability $1 - \alpha$, we bound $\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|^2$ with

$$\tilde{\mathcal{O}} \left(\frac{L\Delta(1 + B_{\text{init}}/\tau)}{T} + \frac{\sigma(\sqrt{L\Delta} + B_{\text{init}} + \sigma)}{\sqrt{TG}} \right),$$

where $\tilde{\mathcal{O}}$ hides constant and logarithmic factors and higher order terms decreasing in T .

Proof. In the case, when $\hat{\beta} = 1$ the worst dependency in (82) w.r.t. T comes from the terms ①, ③, ⑤, and ⑥, and equals $\gamma \sim \frac{1}{T^{1/2}}$. We also have restriction $\gamma \leq \mathcal{O}(1/L)$ and (90). All of those restrictions give the rate of the form

$$\begin{aligned} & \frac{L\Delta}{T} \tilde{\mathcal{O}} \left(1 + \frac{B_{\text{init}}}{\tau} + \frac{\sqrt{B_{\text{init}}}}{\sqrt{\tau}} + \left(\frac{T\sigma^2}{L\Delta G} \right)^{1/2} + \frac{\sigma(\sqrt{L\Delta} + B_{\text{init}} + \sigma)T^{1/2}}{L\Delta\sqrt{G}} \right) \\ &= \tilde{\mathcal{O}} \left(\frac{L\Delta(1 + B_{\text{init}}/\tau)}{T} + \frac{\sigma\sqrt{L\Delta}}{\sqrt{TG}} + \frac{\sigma(\sqrt{L\Delta} + B_{\text{init}} + \sigma)}{\sqrt{TG}} \right) \\ &= \tilde{\mathcal{O}} \left(\frac{L\Delta(1 + B_{\text{init}}/\tau)}{T} + \frac{\sigma(\sqrt{L\Delta} + B_{\text{init}} + \sigma)}{\sqrt{TG}} \right), \end{aligned} \quad (93)$$

where we hide all logarithmic factors in the $\tilde{\mathcal{O}}$ notation. $\square$

E.2 CONVERGENCE IN THE ABSENCE OF BYZANTINE ADVERSARIES

Corollary E.2 (Full statement of Corollary 5.5). *Let Assumptions 3.3, 3.4, and 3.5a and define $B_{\text{init}} := \max\{3\tau, \max_i \{\|\nabla f_i(x^0)\|\} + b\}$. Let the failure probability α be such that $\alpha \in (0, 1)$, and constants $a, \hat{a}, b$, and $\hat{c}$ be defined as in (67), and $\Delta \geq \Phi^0$ for Φ^0 defined in (9). Consider the run of Byz-Clip21-SGD2M (Algorithm 1) for T iterations with DP noise variance σ_ω and $\delta_{\text{byz}} = 0$. Assume the following inequalities hold*

1. step-size restrictions:

(a) $\gamma \leq \frac{1}{24L}$;

(b) $12L\gamma = \beta$;

(c)

$$\frac{5}{6} - \frac{32L^2\beta^2}{\hat{\beta}^2\eta^2}\gamma^2 - \frac{96L^2}{\hat{\beta}^2\eta^2}\gamma^2 \geq 0. \quad (94)$$

2. momentum restrictions:

(a) $\beta \leq \min \left\{ \frac{8\hat{\beta}\tau}{431\sqrt{L\Delta}}, 1 \right\}$;

(b) $\beta \leq \min \left\{ \frac{2\hat{\beta}\tau}{39(B_{\text{init}} - \tau)}, 1 \right\}$;

Algorithm 2 Byz-Clip21-SGD2M without DP adversaries

```
1: Input:  $x^0 \in X$ , momentum parameter  $\beta \in (0, 1]$ , step-size  $\gamma > 0$ 
2: for  $t = 0, \dots, T - 1$  do
3:    $x^{t+1} = x^t - \gamma g^t$ 
4:   for  $i \in \mathcal{G}$  do
5:      $v_i^{t+1} = (1 - \beta)v_i^t + \beta \nabla f_i(x^{t+1}, \xi_i^{t+1})$ 
6:   end for
7:   for  $i \in \mathcal{B}$  do
8:      $v_i^{t+1} = (*)$  sends arbitrary vector
9:   end for
10:   $g^{t+1} = \text{RAgg}(v_1^{t+1}, \dots, v_n^{t+1})$ 
11: end for
```

$$(c) \beta \leq \min \left\{ \frac{2\hat{\beta}\tau}{69b}, 1 \right\};$$

$$(d) \hat{\beta} \leq \min \left\{ \frac{\sqrt{L\Delta}}{a}, 1 \right\};$$

(e) and momentum restrictions defined in (71), (72), (73), (74), (75), (77), (76), (78), (79), and (80) with $\delta_{\text{byz}} = 0$;

Then, with probability $1 - \alpha$, we bound $\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|^2$ with

$$\sqrt{L\Delta}(\sqrt{L\Delta} + B_{\text{init}} + \sigma)\tilde{\mathcal{O}} \left(\frac{\sqrt{d}\sigma_\omega}{\sqrt{GT}\tau} + \frac{d^{1/3}\sigma_\omega^{2/3}}{(TG)^{1/3}\tau^{2/3}} \right),$$

where $\tilde{\mathcal{O}}$ hides constant and logarithmic factors and higher order terms decreasing in T .

Proof. We should simply set $\delta_{\text{byz}} = 0$ in the convergence proof of Theorem 5.1 (Theorem D.1) to obtain the necessary result. $\square$

E.3 CONVERGENCE IN THE ABSENCE OF DP ADVERSARIES

In this setting, server-side momentum and clipping are unnecessary. We thus focus on the specialization of Byz-Clip21-SGD2M with $\hat{\beta} = 1$ and $\tau = +\infty$, shown in Algorithm 2. Under these settings, the method reduces to the algorithm of Karimireddy et al. [2020]. In particular, the error-feedback mechanism is disabled, so the server does not store g_i^t . The analysis correspondingly simplifies and only requires a simpler Lyapunov function

$$\tilde{\Phi}^t := \delta^t + \frac{\gamma}{\beta} \|\bar{v}^t - \nabla f(x^t)\|^2. \quad (95)$$

Also, constants b and c from (67) should be adjusted accordingly. We make use of their replacements $\tilde{b}$ and $\tilde{c}$ defined as

$$\begin{aligned} \tilde{b}^2 &:= 2\sigma^2 \log \left(\frac{8(T+1)G}{\alpha} \right), \quad \tilde{c}^2 := \left(\sqrt{2} + 2\sqrt{3 \log \frac{8(T+1)}{\alpha}} \right)^2 \sigma^2, \\ \tilde{z}^2 &:= \left(\sqrt{2} + 2\sqrt{3 \log \frac{8G(T+1)}{\alpha}} \right)^2 \sigma^2. \end{aligned} \quad (96)$$

E.3.1 Useful Lemmas

We will need to re-derive several descent lemmas as well as introduce new ones.

Lemma E.3. *Let $\{v_i^t\}$ be generated by Algorithm 2. Then we have with probability at least $1 - \frac{\alpha}{8(T+1)}$*

$$\frac{1}{G} \sum_{i \in \mathcal{G}} \|\bar{v}^t - v_i^t\|^2 \leq 2\zeta^2 + 2\beta\tilde{z}^2.$$

Proof. We have

$$\begin{aligned}
\frac{1}{G} \sum_{i \in \mathcal{G}} \|v_i^t - \bar{v}^t\|^2 &\stackrel{(i)}{=} \frac{1}{G} \sum_{i \in \mathcal{G}} \|(1 - \beta)(v_i^{t-1} - \bar{v}^{t-1}) + \beta(\nabla f_i(x^t) - \nabla f(x^t)) + \beta(\theta_i^t - \theta^t)\|^2 \\
&\stackrel{(ii)}{=} \frac{1}{G} \left\| \beta \sum_{k=0}^t (1 - \beta)^{t-k} (\nabla f_i(x^k) - \nabla f(x^k)) + \beta \sum_{k=0}^t (1 - \beta)^{t-k} (\theta_i^k - \theta^k) \right\|^2 \\
&\stackrel{(iii)}{\leq} \frac{2}{G} \sum_{i \in \mathcal{G}} \left\| \beta \sum_{k=0}^t (1 - \beta)^{t-k} (\nabla f_i(x^k) - \nabla f(x^k)) \right\|^2 + \frac{2\beta^2}{G} \left\| \sum_{k=0}^t (1 - \beta)^{t-k} (\theta_i^k - \theta^k) \right\|^2 \\
&\stackrel{(iv)}{\leq} \frac{2(1 - (1 - \beta)^{t+1})}{G} \sum_{i \in \mathcal{G}} \sum_{k=0}^t \beta (1 - \beta)^{t-k} \|\nabla f_i(x^k) - \nabla f(x^k)\|^2 \\
&\quad + \frac{2\beta^2}{G} \sum_{i \in \mathcal{G}} \left\| \sum_{k=0}^t (1 - \beta)^{t-k} (\theta_i^k - \theta^k) \right\|^2, \\
&\stackrel{(v)}{\leq} 2 \sum_{k=0}^t \beta (1 - \beta)^{t-k} \zeta^2 + \frac{2\beta^2}{G} \sum_{i \in \mathcal{G}} \left\| \sum_{k=0}^t (1 - \beta)^{t-k} (\theta_i^k - \theta^k) \right\|^2 \\
&\leq 2\zeta^2 + \frac{2\beta^2}{G} \sum_{i \in \mathcal{G}} \left\| \sum_{k=0}^t (1 - \beta)^{t-k} (\theta_i^k - \theta^k) \right\|^2, \tag{97}
\end{aligned}$$

where (i) – from the update rules of v_i^t and decomposition $\nabla f_i(x^t) + \theta_i^t = \nabla f_i(x^t, \xi_i^t)$, (ii) – unrolling the recursion till the zero iteration and using the initialization with zeros, (iii) – from Young's inequality, (iv) – from Jensen's inequality, (v) – from Assumption 3.5. For the second term in (97), we use Lemma B.2. We know that $\theta_i^k - \theta^k$ is a zero-centered sub-Gaussian r.v. conditioned on all events before iteration k . Therefore, we have

$$\left\| \sum_{k=0}^t (1 - \beta)^{t-k} (\theta_i^k - \theta^k) \right\|^2 \leq (\sqrt{2} + 2u) \sqrt{\sum_{k=0}^t (1 - \beta)^{2(t-k)} \sigma^2} \leq (\sqrt{2} + 2u) \frac{\sigma}{\sqrt{\beta}} \leq \frac{\tilde{z}}{\sqrt{\beta}}, \tag{98}$$

where $u = \sqrt{3 \log \frac{8G(T+1)}{\alpha}}$. Thus, we obtain with probability at least $1 - G \cdot \frac{\alpha}{8G(T+1)}$ that

$$\frac{1}{G} \sum_{i \in \mathcal{G}} \|v_i^t - \bar{v}^t\|^2 \leq 2\zeta^2 + 2\beta\tilde{z}^2. \tag{99}$$

□

Next, we re-derive the descent lemma for the function value.

Lemma E.4. *Let f be L -smooth, Assumption 3.5 hold with $A = 0$, $F^t := f(x^t) - f^*$, $\{x^t\}$ be generated by Algorithm 2. Assume that*

1. $\gamma \leq \frac{1}{2L}$;
2. $\|\theta_i^t\| \leq \tilde{b}$;
3. $\frac{1}{G} \sum_{i \in \mathcal{G}} \|\bar{v}^{t-1} - v_i^{t-1}\|^2 \leq 2\zeta^2 + 2\beta\tilde{z}^2$;

Then we have

$$f(x^{t+1}) \leq f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \frac{1}{4\gamma} \|x^{t+1} - x^t\|^2 + \gamma \|\nabla f(x^t) - \bar{v}^t\|^2 + 2c\delta_{\text{byz}}\zeta^2 + 2c\delta_{\text{byz}}\beta\tilde{z}^2. \tag{100}$$

Proof. Using the derivations from [Islamov et al., 2025b], Lemma 2 we first get

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

We continue the derivations as follows

$$\begin{aligned}
f(x^{t+1}) &\stackrel{(i)}{\leq} f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \frac{\gamma}{4} \|g^t\|^2 + \gamma \|\nabla f(x^t) - \bar{v}^t\|^2 + \gamma \|\bar{v}^t - g^t\|^2 \\
&\stackrel{(ii)}{=} f(x^t) - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \frac{1}{4\gamma} \|x^{t+1} - x^t\|^2 + \gamma \|\nabla f(x^t) - \bar{v}^t\|^2 + \gamma \|\bar{v}^t - g^t\|^2,
\end{aligned} \tag{101}$$

where (i) follows from Jensen's inequality applied to $\|\cdot\|^2$, (ii) — from the update rule of x^t . We bound the term $\|\bar{v}^t - g^t\|^2$ using properties of the aggregator.

$$\begin{aligned}
\|\bar{v}^t - g^t\|^2 &= \|\bar{v}^t - \text{RAgg}(v_1^t, \dots, v_n^t)\|^2 \\
&\leq \frac{c\delta_{\text{byz}}}{G} \sum_{i \in \mathcal{G}} \|v_i^t - \bar{v}^t\|^2 \\
&\stackrel{(iii)}{\leq} 2c\delta_{\text{byz}}\zeta^2 + 2c\delta_{\text{byz}}\beta\tilde{z}^2.
\end{aligned} \tag{102}$$

where (iii) holds by Lemma E.3. Plugging in (102) into (101), we obtain the statement of the lemma. $\square$

Lemma E.5. *Let each f_i be L -smooth, and $\tilde{\Delta} \geq \Phi^0$. Assume that the following inequalities hold for the iterates generated by *Byz-Clip21-SGD2M**

1. $\gamma \leq \frac{1}{6L}$,
2. $\beta \in [0, 1]$,
3. $\|\bar{v}^{t-1}\| \leq \sqrt{64L\tilde{\Delta}} + 3\frac{\tilde{c}}{\sqrt{G}} + 3\sqrt{2c\delta_{\text{byz}}}\zeta + 3\sqrt{\beta}\sqrt{2c\delta_{\text{byz}}}\tilde{z}$,
4. $\|\nabla f(x^{t-1}) - \bar{v}^{t-1}\| \leq \sqrt{4L\tilde{\Delta}} + \frac{3}{2}\frac{\tilde{c}}{\sqrt{G}} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\zeta + \frac{3}{2}\sqrt{\beta}\sqrt{2c\delta_{\text{byz}}}\tilde{z}$,
5. $\|\theta^t\| \leq \frac{\tilde{c}}{\sqrt{G}}$,
6. $\frac{1}{G} \sum_{i \in \mathcal{G}} \|\bar{v}^{t-1} - v_i^{t-1}\|^2 \leq 2c\delta_{\text{byz}}\zeta^2 + 2c\delta_{\text{byz}}\beta\tilde{z}^2$,
7. $\Phi^{t-1} \leq 2\tilde{\Delta}$.

Then we have

$$\|\bar{v}^t\| \leq \sqrt{64L\tilde{\Delta}} + 3\frac{\tilde{c}}{\sqrt{G}} + 3\sqrt{2c\delta_{\text{byz}}}\zeta + 3\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z}. \tag{103}$$

Proof. We start as follows

$$\begin{aligned}
\|\bar{v}^t\| &\stackrel{(i)}{=} \|(1-\beta)\bar{v}^{t-1} + \beta\nabla f(x^t, \xi^t)\| \\
&= \|(1-\beta)(\bar{v}^{t-1} - \nabla f(x^{t-1})) + \beta(\nabla f(x^t) - \nabla f(x^{t-1})) + \beta(\nabla f(x^t, \xi^t) - \nabla f(x^t)) + \nabla f(x^{t-1})\| \\
&\stackrel{(ii)}{\leq} (1-\beta)\|\bar{v}^{t-1} - \nabla f(x^{t-1})\| + \beta L\gamma\|g^{t-1}\| + \beta\|\theta^t\| + \|\nabla f(x^{t-1})\|.
\end{aligned} \tag{104}$$

where (i) follows from the update rule of $\bar{v}^t$, (ii) — from triangle inequality and L -smoothness. We continue the derivations as follows

$$\|\bar{v}^t\| \stackrel{(iii)}{\leq} \sqrt{2L(f(x^{t-1}) - f^*)} + (1-\beta)\|\bar{v}^{t-1} - \nabla f(x^{t-1})\| + \beta\frac{\tilde{c}}{\sqrt{G}} + \beta L\gamma\|\bar{v}^{t-1}\| + \beta L\gamma\|\bar{v}^{t-1} - g^{t-1}\|. \tag{105}$$

where (iii) follows from L -smoothness, assumption 5 of the lemma, and triangle inequality. The term $\|\bar{v}^{t-1} - g^{t-1}\|$ can be

bounded by $\sqrt{2c\delta_{\text{byz}}\zeta^2 + 2c\delta_{\text{byz}}\beta\tilde{c}^2}$ by (102) and assumptions 5, 6 of the lemma. We continue further bounding as follows

$$\begin{aligned}
\|\bar{v}^t\| &\stackrel{(iv)}{\leq} \sqrt{4L\tilde{\Delta}} + (1-\beta) \left(\sqrt{4L\tilde{\Delta}} + \frac{3}{2} \frac{\tilde{c}}{\sqrt{G}} + \frac{3}{2} \sqrt{2c\delta_{\text{byz}}}\zeta + \frac{3}{2} \sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z} \right) + \beta \frac{\tilde{c}}{\sqrt{G}} \\
&\quad + L\gamma\beta \left(\sqrt{64L\tilde{\Delta}} + 3 \frac{\tilde{c}}{\sqrt{G}} + 3\sqrt{2c\delta_{\text{byz}}}\zeta + 3\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z} \right) + \beta L\gamma\sqrt{2c\delta_{\text{byz}}}\zeta + \beta L\gamma\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z} \\
&= \sqrt{L\tilde{\Delta}}(2 + 2(1-\beta) + 8L\gamma\beta) + \frac{\tilde{c}}{\sqrt{G}}(3/2(1-\beta) + \beta + 3L\gamma\beta) + \sqrt{2c\delta_{\text{byz}}}\zeta(3/2(1-\beta) + 3L\gamma\beta + \beta L\gamma) \\
&\quad + \sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z}(3/2(1-\beta) + 3L\gamma\beta + \beta L\gamma). \tag{106}
\end{aligned}$$

For the first coefficient preceding $\tilde{\Delta}$, we have

$$2 + 2(1-\beta) + 8L\gamma\beta \leq 8 \Leftarrow 1-\beta + 4L\gamma\beta \leq 3 \leq L\gamma\beta \leq 1,$$

where the last inequality holds by the choice of β and γ . For the second coefficient preceding $\frac{\tilde{c}}{\sqrt{G}}$, we have

$$\frac{3}{2}(1-\beta) + \beta + 3L\gamma\beta \leq 3 \Leftarrow 4L\gamma\beta \leq \frac{3}{2}(1+\beta),$$

where the last inequality holds by the choice of β and γ . For the third coefficient preceding $\sqrt{2c\delta_{\text{byz}}}\zeta$, we have

$$\frac{3}{2}(1-\beta) + 3L\gamma\beta + \beta L\gamma \leq 3 \Leftarrow 3L\gamma\beta \leq 1,$$

where the last inequality holds by the choice of β and γ . For the fourth coefficient preceding $\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z}$, we have

$$\frac{3}{2}(1-\beta) + 3L\gamma\beta + \beta L\gamma \leq 3 \Leftarrow 4L\gamma\beta \leq \frac{3}{2}(1+\beta),$$

where the last inequality holds by the choice of β and γ . □

Lemma E.6. Let each f_i be L -smooth, $\tilde{\Delta} \geq \Phi^0$. Assume that the following inequalities hold for the iterates generated by *Byz-Clip21-SGD2M*

1. $\gamma \leq \frac{1}{6L}$;
2. $4L\gamma \leq \beta$;
3. $\beta \in [0, 1]$;
4. $\frac{1}{G} \sum_{i \in \mathcal{G}} \|v_i^{t-1} - \bar{v}^{t-1}\|^2 \leq 2c\delta_{\text{byz}}\zeta^2 + 2c\delta_{\text{byz}}\beta\tilde{z}^2$;
5. $\|\bar{v}^{t-1}\| \leq \sqrt{64L\tilde{\Delta}} + 3\frac{\tilde{c}}{\sqrt{G}} + 3\sqrt{2c\delta_{\text{byz}}}\zeta + 3\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z}$,
6. $\|\nabla f(x^{t-1}) - \bar{v}^{t-1}\| \leq \sqrt{4L\tilde{\Delta}} + \frac{3}{2}\frac{\tilde{c}}{\sqrt{G}} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\zeta + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z}$;
7. $\|\theta^t\| \leq \frac{\tilde{c}}{\sqrt{G}}$;
8. $\|\theta_i^t\| \leq \tilde{b}$ for all $i \in \mathcal{G}$;
9. $\Phi^{t-1} \leq 2\tilde{\Delta}$.

Then we have

$$\|\nabla f(x^t) - \bar{v}^t\| \leq \sqrt{4L\tilde{\Delta}} + \frac{3}{2} \frac{\tilde{c}}{\sqrt{G}} + \frac{3}{2} \sqrt{2c\delta_{\text{byz}}}\zeta + \frac{3}{2} \sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z}. \tag{107}$$

Proof. We have

$$\begin{aligned}
\|\nabla f(x^t) - v^t\| &\stackrel{(i)}{=} \|\nabla f(x^t) - (1 - \beta)\bar{v}^{t-1} - \beta\nabla f_i(x^t, \xi^t)\| \\
&\stackrel{(ii)}{\leq} (1 - \beta)\|\nabla f(x^t) - \bar{v}^{t-1}\| + \beta\|\nabla f(x^t) - \nabla f(x^t, \xi^t)\| \\
&\stackrel{(iii)}{\leq} (1 - \beta)\|\nabla f(x^t) - \nabla f(x^{t-1})\| + (1 - \beta)\|\nabla f(x^{t-1}) - \bar{v}^{t-1}\| + \beta\|\theta^t\| \\
&\stackrel{(iv)}{\leq} (1 - \beta)L\gamma\|g^{t-1}\| + (1 - \beta)\|\nabla f(x^{t-1}) - \bar{v}^{t-1}\| + \beta\|\theta^t\| \\
&\stackrel{(v)}{\leq} (1 - \beta)L\gamma\|\bar{v}^{t-1} - g^{t-1}\| + (1 - \beta)L\gamma\|\bar{v}^{t-1}\| + (1 - \beta)\|\nabla f(x^{t-1}) - \bar{v}^{t-1}\| + \beta\|\theta^t\|, \quad (108)
\end{aligned}$$

where (i) follows from the update rule of v^t , (ii)-(iii) — from the triangle inequality, (iv) — from the update rule of x^t and L -smoothness, (v) — from triangle inequality. We continue as follows

$$\begin{aligned}
\|\nabla f(x^t) - \bar{v}^t\| &\stackrel{(vi)}{\leq} \sqrt{2c\delta_{\text{byz}}}(1 - \beta)L\gamma\zeta + \sqrt{2c\delta_{\text{byz}}}(1 - \beta)L\gamma\sqrt{\beta}\tilde{z} + (1 - \beta)L\gamma\|\bar{v}^{t-1}\| + \beta\|\theta^t\| \\
&\quad + (1 - \beta)\|\nabla f(x^{t-1}) - \bar{v}^{t-1}\| \\
&\stackrel{(vii)}{\leq} \sqrt{2c\delta_{\text{byz}}}(1 - \beta)L\gamma\zeta + \sqrt{2c\delta_{\text{byz}}}(1 - \beta)L\gamma\sqrt{\beta}\tilde{z} \\
&\quad + (1 - \beta)L\gamma\left(\sqrt{64L\tilde{\Delta}} + 3\frac{\tilde{c}}{\sqrt{G}} + 3\sqrt{2c\delta_{\text{byz}}}\zeta + 3\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z}\right) \\
&\quad + \beta\frac{\tilde{c}}{\sqrt{G}} + (1 - \beta)\left(\sqrt{4L\tilde{\Delta}} + \frac{3}{2}\frac{\tilde{c}}{\sqrt{G}} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\zeta + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z}\right) \\
&= \sqrt{L\tilde{\Delta}}(8(1 - \beta)L\gamma + 2(1 - \beta)) + \frac{\tilde{c}}{\sqrt{G}}(3(1 - \beta)L\gamma + 3/2(1 - \beta)) \\
&\quad + \sqrt{2c\delta_{\text{byz}}}\zeta((1 - \beta)L\gamma + 3(1 - \beta)L\gamma + 3/2(1 - \beta)) \\
&\quad + \sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z}((1 - \beta)L\gamma + 3(1 - \beta)L\gamma + 3/2(1 - \beta)).
\end{aligned}$$

where (vi) follows from (102) and assumptions 4, 8 of the lemma, (vii) — from assumptions 5, 6, 7 of the lemma.

For the first coefficient preceding $\sqrt{L\tilde{\Delta}}$, we have

$$8(1 - \beta)L\gamma + 2(1 - \beta) \leq 2 \Leftrightarrow 4L\gamma \leq \beta,$$

where the last inequality holds by assumption 2 of the lemma. For the second coefficient preceding $\frac{\tilde{c}}{\sqrt{G}}$, we have

$$3(1 - \beta)L\gamma + \frac{3}{2}(1 - \beta) \leq \frac{3}{2} \Leftrightarrow 2L\gamma \leq \beta,$$

where the last inequality holds by assumption 2 of the lemma. For the third coefficient preceding $\sqrt{2c\delta_{\text{byz}}}\zeta$, we have

$$4(1 - \beta)L\gamma + \frac{3}{2}(1 - \beta) \leq \frac{3}{2} \Leftrightarrow 4L\gamma \leq \frac{3}{2}\beta \Leftrightarrow \frac{8}{3}L\gamma \leq \beta,$$

where the last inequality holds by assumption 2 of the lemma. Finally, for the fourth coefficient preceding $\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z}$, we have

$$4(1 - \beta)L\gamma + \frac{3}{2}(1 - \beta) \leq \frac{3}{2} \Leftrightarrow 4L\gamma \leq \frac{3}{2}\beta \Leftrightarrow \frac{8}{3}L\gamma \leq \beta,$$

where the last inequality holds by assumption 2 of the lemma. This finalizes the proof. $\square$

E.3.2 Main Convergence Theorem

Theorem E.7 (Full statement of Theorem 5.6). *Let Assumptions 3.3, 3.4, and 3.5 hold. Let the failure probability $\alpha \in (0, 1)$, and constants $\tilde{b}$ and $\tilde{c}$ be defined as in (96), and $\tilde{\Delta} \geq \tilde{\Phi}^0$ for $\tilde{\Phi}^0$ defined in (95). Consider the run of Byz-Clip21-SGD2M with $\sigma_\omega = 0$ and $\tau = +\infty$ (Algorithm 2) for T iterations. Assume that the following inequalities are satisfied*

1. $\gamma \leq \frac{1}{6L}$,
2. $\beta \geq 4L\gamma$,
3. and momentum restrictions defined in (112), (113), and (114).

Then, with probability at least $1 - \alpha$, we bound $\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|^2$ with

$$\tilde{O} \left(\frac{L\tilde{\Delta}}{T} + \left(\frac{\sigma^2 L\tilde{\Delta}}{GT} \right)^{1/2} + \frac{\sigma(\sqrt{L\tilde{\Delta}} + \sigma/\sqrt{G} + \sqrt{c\delta_{\text{byz}}}\zeta + \sqrt{c\delta_{\text{byz}}}\sigma)}{\sqrt{GT}} + c\delta_{\text{byz}}\zeta^2 \right), \quad (109)$$

where we can choose $\tilde{\Delta} = 2(f(x^0) - f^*) = 2F^0$.

Proof. For convenience, we define $\nabla f_i(x^{-1}, \xi_i^{-1}) = v_i^{-1} = g_i^{-1} = 0, \Phi^{-1} = \Phi^0$. Next, let us define an event $\tilde{E}^t$ for each $t \in \{0, \dots, T\}$ such that the following inequalities hold for all $k \in \{0, \dots, t\}$

1. $\frac{1}{G} \sum_{i \in \mathcal{G}} \|\bar{v}^t - v_i^t\| \leq 2\zeta^2 + 2\beta\tilde{z}^2$;
2. $\|\theta_i^k\| \leq b$ for all $i \in \mathcal{G}$ and $\|\theta^k\| \leq \frac{\tilde{c}}{\sqrt{G}}$;
3. $\|\bar{v}^{t-1}\| \leq \sqrt{64L\tilde{\Delta}} + 3\frac{\tilde{c}}{\sqrt{G}} + 3\sqrt{2c\delta_{\text{byz}}}\zeta + 3\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z}$,
4. $\|\nabla f(x^{t-1}) - \bar{v}^{t-1}\| \leq \sqrt{4L\tilde{\Delta}} + \frac{3}{2}\frac{\tilde{c}}{\sqrt{G}} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\zeta + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z}$;
5. $\tilde{\Phi}^k \leq 2\tilde{\Delta}$;
- 6.

$$\frac{1}{2}\Delta \geq 2\gamma(1 - \beta) \sum_{l=0}^{k-1} \langle \nabla f(x^l) - \nabla f(x^{l+1}), \theta^{l+1} \rangle.$$

Then, we will derive the result by induction, i.e., using the induction w.r.t. t , we will show that $\mathbb{P}(\tilde{E}^t) \geq 1 - \frac{\alpha(t+1)}{T+1}$ for all $t \in \{0, \dots, T-1\}$.

Before moving on to the proof's induction part, we need to establish several useful bounds. Denote the events $\tilde{\Theta}_i^t$ and $\tilde{\Theta}^t$ as

$$\tilde{\Theta}_i^t := \{\|\theta_i^t\| \geq \tilde{b}\}, \quad \tilde{\Theta}^t := \left\{ \|\theta^t\| \geq \frac{\tilde{c}}{\sqrt{G}} \right\}, \quad \hat{\Theta}_i^t := \left\{ \left\| \sum_{k=0}^t (1 - \beta)^{t-k} (\theta_i^k - \theta^k) \right\| \geq \frac{\tilde{z}}{\sqrt{\beta}} \right\} \quad (110)$$

respectively. From Assumption 3.4 we have

$$\mathbb{P}(\tilde{\Theta}_i^t) \leq 2 \exp \left(-\frac{\tilde{b}^2}{2\sigma^2} \right) = \frac{\alpha}{8(T+1)G}$$

where the last equality is by definition of $\tilde{b}^2$. Therefore, $\mathbb{P}(\tilde{\Theta}_i^t) \geq 1 - \frac{\alpha}{8G(T+1)}$. Besides, notice that the constant $\tilde{c}$ in (96) can be viewed as

$$\tilde{c}^2 = (\sqrt{2} + 2\tilde{b}_3)\sigma^2 \quad \text{where} \quad \tilde{b}_3^2 = 3 \log \frac{8(T+1)}{\alpha}.$$

Now, we can use Lemma B.2 to bound $\mathbb{P}(\tilde{\Theta}^t)$. Since all θ_i^t are independent σ -sub-Gaussian random vectors, then we have

$$\mathbb{P} \left(\left\| \sum_{i \in \mathcal{G}} \theta_i^t \right\| \geq \tilde{c}\sqrt{G} \right) = \mathbb{P} \left(\|\theta^t\| \geq \frac{\tilde{c}}{\sqrt{G}} \right) \leq \exp(-\tilde{b}_3^2/3) = \frac{\alpha}{8(T+1)}.$$

The bound $\mathbb{P}(\hat{\Theta}_i^t) \leq \frac{\alpha}{6G(T+1)}$ was shown before in Lemma E.3.

Now, we are ready to prove that $\mathbb{P}(\tilde{E}^t) \geq 1 - \frac{\alpha(t+1)}{T+1}$ for all $t \in \{0, \dots, T-1\}$. First, we show that the base of induction holds.

Base of induction.

1. To establish the first, we use Lemma E.3, which guarantees the bound in $\cap_{i \in \mathcal{G}} \bar{\Theta}^t$, i.e., with probability $1 - \frac{\alpha}{8(T+1)}$.
2. $\bar{v}^0 = \frac{1}{G} \sum_{i \in \mathcal{G}} ((1 - \beta)v_i^{-1} + \beta \nabla f_i(x^0, \xi_i^0)) = \frac{\beta}{G} \nabla f(x^0, \xi^0)$. Therefore, we have

$$\begin{aligned}
\|\bar{g}^0\| &= \beta \|\nabla f(x^0, \xi^0)\| \\
&\leq \beta \|\nabla f(x^0)\| + \beta \|\nabla f(x^0, \xi^0) - \nabla f(x^0)\| \\
&\leq \beta \sqrt{2L(f(x^0) - f^*)} + \beta \frac{\tilde{c}}{\sqrt{G}} \\
&\leq \beta \sqrt{2L\tilde{\Phi}^0} + \beta \frac{\tilde{c}}{\sqrt{G}} \\
&\leq \beta \sqrt{4L\tilde{\Delta}} + \beta \frac{\tilde{c}}{\sqrt{G}} \\
&\leq \sqrt{64L\tilde{\Delta}} + \frac{3\tilde{c}}{\sqrt{G}} + 3\sqrt{2c\delta_{\text{byz}}}\zeta + 3\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z}.
\end{aligned}$$

The inequalities above again hold in $\bar{\Theta}^0$, i.e., with probability at least $1 - \frac{\alpha}{8(T+1)}$. Therefore, the condition 3 of the induction is verified.

3. We have

$$\begin{aligned}
\|\bar{v}^0 - \nabla f(x^0)\| &= \|\beta \nabla f(x^0, \xi^0) - \nabla f(x^0)\| \\
&\leq \beta \|\nabla f(x^0, \xi^0) - \nabla f(x^0)\| + (1 - \beta) \|\nabla f(x^0)\| \\
&\leq \beta b + (1 - \beta) \sqrt{4L\tilde{\Delta}}.
\end{aligned}$$

The bound above holds with probability at least $1 - \frac{\alpha}{8(T+1)}$ because it holds in $\cap_{i \in \mathcal{G}} \bar{\Theta}_i^0$. Therefore, the bound 4 of the assumption of the induction is verified.

4. Next, we emphasize that the condition 8 of the induction assumption also hold, as $\tilde{\Phi}^0 \leq 2\tilde{\Phi}^0 \leq 2\tilde{\Delta}$ by the choice of $\tilde{\Delta}$.
5. We finalize the induction base by noting that the condition 6 of the induction assumption holds since the RHS equals 0.

Therefore, we conclude that the conditions 1-8 hold with a probability of at least

$$\begin{aligned}
\mathbb{P}(\bar{\Theta}^0 \cap (\cap_{i \in \mathcal{G}} \bar{\Theta}_i^0) \cap (\cap_{i \in \mathcal{G}} \bar{\Theta}_i^0)) &\geq 1 - \mathbb{P}(\bar{\Theta}^0) - \sum_{i \in \mathcal{G}} \mathbb{P}(\bar{\Theta}_i^0) - \sum_{i \in \mathcal{G}} \mathbb{P}(\bar{\Theta}_i^0) \\
&\geq 1 - \frac{\alpha}{8(T+1)} - G \cdot \frac{\alpha}{8G(T+1)} - G \cdot \frac{\alpha}{8G(T+1)} \\
&= 1 - \frac{\alpha}{2(T+1)} > 1 - \frac{\alpha}{T+1},
\end{aligned}$$

i.e., $\mathbb{P}(E^0) \geq 1 - \frac{\alpha}{T+1}$ holds. This is the base of the induction.

Transition step of induction. Assume that all events $\bar{\Theta}^{K+1}, \bar{\Theta}_i^{K+1}$ take place, i.e., $\|\theta_i^{K+1}\| \leq \tilde{b}, \|\theta^{K+1}\| \leq \frac{\tilde{c}}{\sqrt{G}}$ for all $i \in \mathcal{G}$. That is, we assume that the event

$$\bar{\Theta}^{K+1} \cap (\cap_{i \in \mathcal{G}} \bar{\Theta}_i^{K+1}) \cap (\cap_{i \in \mathcal{G}} \bar{\Theta}_i^{K+1}) \cap \tilde{E}^K$$

holds. Then, by the assumptions of the induction and from Lemma E.5, we get that

$$\|\bar{v}^{K+1}\| \leq \sqrt{64L\tilde{\Delta}} + 3\frac{\tilde{c}}{\sqrt{G}} + 3\sqrt{2c\delta_{\text{byz}}}\zeta + 3\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z},$$

from Lemma E.6 we get that

$$\|\nabla f(x^{K+1}) - v^{K+1}\| \leq \sqrt{4L\tilde{\Delta}} + \frac{3}{2}\frac{\tilde{c}}{\sqrt{G}} + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\zeta + \frac{3}{2}\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z},$$

from Lemma E.5 we get

$$\frac{1}{G} \sum_{i \in \mathcal{G}} \|\bar{v}^{K+1} - v_i^{K+1}\|^2 \leq 2\zeta^2 + 2\beta\bar{z}^2.$$

This means that conditions 1-5 in the induction assumption are also verified for the step $K+1$. Since for all $t \in \{0, \dots, K+1\}$ inequalities 1-7 are verified, we can write for each $t \in \{0, \dots, K\}$ by Lemmas C.7 and E.4 the following

$$\begin{aligned} \tilde{\Phi}^{t+1} &= \delta^{t+1} + \frac{\gamma}{\beta} \tilde{P}^{t+1} \\ &\leq \delta^t - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 - \frac{1}{4\gamma} R^t + \gamma \tilde{P}^t + 2c\delta_{\text{byz}}\gamma(\zeta^2 + \beta\bar{z}^2) \\ &\quad + \frac{\gamma}{\beta} \left((1-\beta)\tilde{P}^t + \frac{3L^2}{\beta} R^t + \beta^2 \frac{\tilde{c}^2}{G} + 2\beta(1-\beta)\langle \bar{v}^t - \nabla f(x^{t+1}), \theta^{t+1} \rangle \right). \end{aligned}$$

Rearranging terms, we get

$$\begin{aligned} \tilde{\Phi}^{t+1} &\leq \delta^t - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 + \frac{\gamma}{\beta} \tilde{P}^t (\beta + 1 - \beta) + 2c\delta_{\text{byz}}\gamma(\zeta^2 + \beta\bar{z}^2) - \frac{1}{4\gamma} R^t \left(1 - \frac{12L^2}{\beta^2} \gamma^2 \right) + \tilde{c}^2 \frac{\gamma\beta}{G} \\ &\quad + 2\gamma(1-\beta)\langle \nabla f(x^t) - \nabla f(x^{t+1}), \theta^{t+1} \rangle. \end{aligned}$$

Using step-size restriction (ii), we get rid of the term with R^t and obtain

$$\tilde{\Phi}^{t+1} \leq \delta^t - \frac{\gamma}{2} \|\nabla f(x^t)\|^2 + \frac{\gamma}{\beta} \tilde{P}^t + 2c\delta_{\text{byz}}\gamma(\zeta^2 + \beta\bar{z}^2) + \tilde{c}^2 \frac{\gamma\beta}{G} + 2\gamma(1-\beta)\langle \nabla f(x^t) - \nabla f(x^{t+1}), \theta^{t+1} \rangle.$$

Now we sum all the inequalities above for $t \in \{0, \dots, K\}$ and get

$$\tilde{\Phi}^{K+1} \leq \tilde{\Phi}^0 - \frac{\gamma}{2} \sum_{t=0}^K \|\nabla f(x^t)\|^2 + K\tilde{c}^2 \frac{\gamma\beta}{G} + 2c\delta_{\text{byz}}\gamma(\zeta^2 + \beta\bar{z}^2)K + 2\gamma(1-\beta) \sum_{t=0}^K \langle \nabla f(x^t) - \nabla f(x^{t+1}), \theta^{t+1} \rangle. \quad (111)$$

Rearranging terms, we get

$$\frac{\gamma}{2} \sum_{t=0}^K \|\nabla f(x^t)\|^2 \leq \tilde{\Phi}^0 - \tilde{\Phi}^{K+1} + K\tilde{c}^2 \frac{\gamma\beta}{G} + 2c\delta_{\text{byz}}\gamma(\zeta^2 + \beta\bar{z}^2)K + 2\gamma(1-\beta) \sum_{t=0}^K \langle \nabla f(x^t) - \nabla f(x^{t+1}), \theta^{t+1} \rangle.$$

Taking into account that $\frac{\gamma}{2} \sum_{t=0}^K \|\nabla f(x^t)\|^2 \geq 0$, we get that the event $\tilde{E}^K \cap \left(\cap_{i \in \mathcal{G}} \widetilde{\Theta}_i^{K+1} \right) \cap \left(\cap_{i \in \mathcal{G}} \widetilde{\Theta}_i^t \right) \cap \widetilde{\Theta}^{K+1}$ implies

$$\tilde{\Phi}^{K+1} \leq \tilde{\Phi}^0 + K\tilde{c}^2 \frac{\gamma\beta}{G} + 2c\delta_{\text{byz}}\gamma(\zeta^2 + \beta\bar{z}^2)K + 2\gamma(1-\beta) \sum_{t=0}^K \langle \nabla f(x^t) - \nabla f(x^{t+1}), \theta^{t+1} \rangle.$$

Next, we define the following random vector:

$$\zeta_5^t := \begin{cases} \nabla f(x^t) - \nabla f(x^{t+1}), & \text{if } \|\nabla f(x^t) - \nabla f(x^{t+1})\| \leq L\gamma \left(\sqrt{64L\bar{\Delta}} + 3\frac{\tilde{c}}{\sqrt{G}} + 3\sqrt{2c\delta_{\text{byz}}}\zeta + 3\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\bar{z} \right) \\ 0, & \text{otherwise} \end{cases}.$$

By definition, the introduced random vector ζ_5^t is bounded with probability 1. Moreover, by the definition of $\tilde{E}^t$ we get that the event $\tilde{E}^K \cap \widetilde{\Theta}^{K+1} \cap \left(\cap_{i \in \mathcal{G}} \widetilde{\Theta}_i^{K+1} \right) \cap \left(\cap_{i \in \mathcal{G}} \widetilde{\Theta}_i^t \right)$ implies

$$\zeta_5^t = \nabla f(x^t) - \nabla f(x^{t+1}).$$

Therefore, the event $\tilde{E}^K \cap \widetilde{\Theta}^{K+1} \cap \left(\cap_{i \in \mathcal{G}} \widetilde{\Theta}_i^{K+1} \right) \cap \left(\cap_{i \in \mathcal{G}} \widetilde{\Theta}_i^t \right)$ implies

$$\tilde{\Phi}^{K+1} \leq \underbrace{\tilde{\Phi}^0}_{I} + \underbrace{K\tilde{c}^2 \frac{\gamma\beta}{G}}_{II} + \underbrace{\frac{2\gamma(1-\beta)}{G} \sum_{t=0}^K \langle \zeta_5^t, \theta^{t+1} \rangle}_{III} + \underbrace{2c\delta_{\text{byz}}\gamma(\zeta^2 + \beta\bar{z}^2)K}_{III}.$$

Bound of the term I. Since $4L\gamma \leq \beta$, for the term I we have

$$K\tilde{c}^2 \frac{\gamma\beta}{G} \leq K\tilde{c}^2 \frac{\beta^2}{4LG}.$$

By choosing β such that

$$\beta \leq \left(\frac{4L\tilde{\Delta}G}{3T\tilde{c}^2} \right)^{1/2}, \quad (112)$$

we get that

$$K\tilde{c}^2 \frac{\gamma\beta}{G} \leq \frac{\tilde{\Delta}}{2}.$$

This bound holds with probability 1. Note that the worst dependency in the restriction on β w.r.t. T is $\mathcal{O}(1/T^{1/2})$.

Bound of the term III. Since $4L\gamma \leq \beta$, for the term III we have

$$2c\delta_{\text{byz}}\gamma(\zeta^2 + \beta\tilde{z}^2)K \leq \frac{1}{2L}c\delta_{\text{byz}}\beta(\zeta^2 + \beta\tilde{z}^2)K.$$

By choosing β such that

$$\beta \leq \min \left\{ \frac{L\tilde{\Delta}}{3c\delta_{\text{byz}}\zeta^2 T}, \left(\frac{L\tilde{\Delta}}{3c\delta_{\text{byz}}T\tilde{z}^2} \right)^{1/2} \right\}, \quad (113)$$

we get that

$$\frac{1}{2L}c\delta_{\text{byz}}\beta(\zeta^2 + \beta\tilde{z}^2)K \leq \frac{\tilde{\Delta}}{3}.$$

This bound holds with probability 1. Note that the worst dependency in the restriction on β w.r.t. T is $\mathcal{O}(1/T)$.

Bound of the term II. The bound in this case is obtained by using concentration inequality Lemma B.2. We define

$$\sigma_8^2 := 4L^2\gamma^2 \cdot \left(\sqrt{64L\tilde{\Delta}} + 3\frac{\tilde{c}}{\sqrt{G}} + 3\sqrt{2c\delta_{\text{byz}}}\zeta + 3\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z} \right)^2 \cdot \frac{\sigma^2}{G}.$$

Then we have

$$\begin{aligned} & \mathbb{E} \left[\exp \left(\left| \frac{4L^2\gamma^2}{\sigma_8^2} 4(1-\beta)^2 \langle \zeta_5^l, \theta^{l+1} \rangle^2 \right| \right) \mid l \right] \\ & \leq \mathbb{E} \left[\exp \left(\frac{4L^2\gamma^2}{\sigma_8^2} \|\zeta_5^l\|^2 \cdot \|\theta^{l+1}\|^2 \right) \mid l \right] \\ & \leq \mathbb{E} \left[\exp \left(\frac{4L^2\gamma^2}{\sigma_8^2} \left(\sqrt{64L\tilde{\Delta}} + 3\frac{\tilde{c}}{\sqrt{G}} + 3\sqrt{2c\delta_{\text{byz}}}\zeta + 3\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z} \right) \cdot \|\theta^{l+1}\|^2 \right) \mid l \right]. \end{aligned}$$

Since θ^{l+1} is sub-Gaussian with parameter $\frac{\sigma^2}{G}$, then we can continue the chain of inequalities above using the definition of σ_8^2

$$\begin{aligned} & \mathbb{E} \left[\exp \left(\left[4L^2\gamma^2 \cdot \left(\sqrt{64L\tilde{\Delta}} + 3\frac{\tilde{c}}{\sqrt{G}} + 3\sqrt{2c\delta_{\text{byz}}}\zeta + 3\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z} \right)^2 \cdot \frac{\sigma^2}{G} \right]^{-1} \right. \right. \\ & \quad \left. \left. 4L^2\gamma^2 \cdot \left(\sqrt{64L\tilde{\Delta}} + 3\frac{\tilde{c}}{\sqrt{G}} + 3\sqrt{2c\delta_{\text{byz}}}\zeta + 3\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z} \right)^2 \cdot \|\theta^{l+1}\|^2 \right) \mid l \right] \\ & = \mathbb{E} \left[\exp \left(\frac{\|\theta^{l+1}\|^2}{\sigma^2/G} \right) \right] \leq \exp(1). \end{aligned}$$

Therefore, we have by Lemma B.2 that

$$\begin{aligned}
& \Pr \left[2\gamma(1-\beta) \left\| \sum_{t=0}^K \langle \zeta_5^t, \theta^{t+1} \rangle \right\| \right] \\
& \geq (\sqrt{2} + \sqrt{2}b_1) \sqrt{\sum_{t=0}^K 4L^2\gamma^2 \frac{\sigma^2}{G} \left(\sqrt{64L\tilde{\Delta}} + 3\frac{\tilde{c}}{\sqrt{G}} + 3\sqrt{2c\delta_{\text{byz}}}\zeta + 3\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z} \right)^2} \\
& \leq \exp(-\tilde{b}_1^2/3) = \frac{\alpha}{6(T+1)}, \quad \text{where } \tilde{b}_1^2 := 3 \log \left(\frac{6(T+1)}{\alpha} \right)
\end{aligned}$$

Using the restrictions $4L\gamma \leq \beta$, we get

$$\begin{aligned}
& (\sqrt{2} + \sqrt{2}b_1) \sqrt{(K+1)} \cdot \frac{2L\gamma}{\sqrt{G}} \sigma \left(\sqrt{64L\tilde{\Delta}} + 3\frac{\tilde{c}}{\sqrt{G}} + 3\sqrt{2c\delta_{\text{byz}}}\zeta + 3\sqrt{2c\delta_{\text{byz}}}\sqrt{\beta}\tilde{z} \right) \\
& \leq (\sqrt{2} + \sqrt{2}b_1) \sqrt{(K+1)} \cdot \frac{\beta\sigma}{2L\sqrt{G}} \left(\sqrt{64L\tilde{\Delta}} + 3\frac{\tilde{c}}{\sqrt{G}} + 3\sqrt{2c\delta_{\text{byz}}}\zeta + 3\sqrt{2c\delta_{\text{byz}}}\tilde{z} \right) \\
& \leq \frac{\tilde{\Delta}}{3}
\end{aligned}$$

because we choose β

$$\beta \leq \left(\frac{2L\tilde{\Delta}\sqrt{G}}{3\sqrt{2}(1+\tilde{b}_1)\sigma\sqrt{T} \left(\sqrt{64L\tilde{\Delta}} + 3\frac{\tilde{c}}{\sqrt{G}} + 3\sqrt{2c\delta_{\text{byz}}}\zeta + 3\sqrt{2c\delta_{\text{byz}}}\tilde{z} \right)} \right), \quad (114)$$

and $K+1 \leq T$.

This implies

$$\mathbb{P} \left(2\gamma(1-\beta) \left\| \sum_{t=0}^K \langle \zeta_{5,i}^t, \theta^{t+1} \rangle \right\| \geq \frac{\tilde{\Delta}}{3} \right) \leq \frac{\alpha}{6(T+1)}.$$

Note that the worst dependency in the choice of β w.r.t T is $\tilde{O}(1/T^{1/2})$.

Final probability. Therefore, the probability event

$$\Omega := \tilde{E}^K \cap \overline{\tilde{\Theta}^{K+1}} \cap \left(\cap_{i \in \mathcal{G}} \overline{\tilde{\Theta}_i^{K+1}} \right) \cap \left(\cap_{i \in \mathcal{G}} \overline{\hat{\Theta}_i^{K+1}} \right) \cap E_I \cap E_{II} \cap E_{III},$$

where E_I - E_{III} denotes that each of the terms I , II , and III is smaller than $\frac{\tilde{\Delta}}{3}$. This implies that

$$I + II + III \leq \tilde{\Delta},$$

i.e., condition 6 in the induction assumption holds. Moreover, this also implies that

$$\tilde{\Phi}^{K+1} \leq \tilde{\Phi}^0 + \tilde{\Delta} \leq \tilde{\Delta} + \tilde{\Delta} = 2\tilde{\Delta},$$

i.e., condition 5 in the induction assumption holds. The probability $\mathbb{P}(\tilde{E}_{K+1})$ can be lower bounded as follows

$$\begin{aligned}
\mathbb{P}(\tilde{E}_{K+1}) & \geq \mathbb{P}(\Omega) \\
& = \mathbb{P} \left(\tilde{E}_K \cap \overline{\tilde{\Theta}^{K+1}} \cap \left(\cap_{i \in \mathcal{G}} \overline{\tilde{\Theta}_i^{K+1}} \right) \cap \left(\cap_{i \in \mathcal{G}} \overline{\hat{\Theta}_i^{K+1}} \right) \cap E_I \cap E_{II} \cap E_{III} \right) \\
& = 1 - \mathbb{P} \left(\tilde{E}_K \cup \tilde{\Theta}^{K+1} \cup \left(\cup_{i \in \mathcal{G}} \tilde{\Theta}_i^{K+1} \right) \cup \left(\cup_{i \in \mathcal{G}} \hat{\Theta}_i^{K+1} \right) \cup \overline{E}_I \cup \overline{E}_{II} \cup \overline{E}_{III} \right) \\
& \geq 1 - \mathbb{P}(\tilde{E}_K) - \mathbb{P}(\tilde{\Theta}^{K+1}) - \sum_{i \in \mathcal{G}} \mathbb{P}(\tilde{\Theta}_i^{K+1}) - \sum_{i \in \mathcal{G}} \mathbb{P}(\hat{\Theta}_i^{K+1}) - \underbrace{\mathbb{P}(\overline{E}_I)}_{=0} - \underbrace{\mathbb{P}(\overline{E}_{II})}_{=0} - \underbrace{\mathbb{P}(\overline{E}_{III})}_{=0} \\
& \geq 1 - \frac{\alpha(K+1)}{T+1} - \frac{\alpha}{8(T+1)} - \sum_{i \in \mathcal{G}} \frac{\alpha}{8G(T+1)} - \sum_{i \in \mathcal{G}} \frac{\alpha}{8G(T+1)} - \frac{\alpha}{8(T+1)} \\
& \geq 1 - \frac{\alpha(K+2)}{T+1}.
\end{aligned}$$

Algorithm 3 Byz-Clip-SGD

```

1: Input:  $x^0 \in X$ , step-size  $\gamma > 0$ , clipping parameter
    $\tau > 0$ , DP noise variance  $\sigma_\omega^2 \geq 0$ 
2:
3: for  $t = 0, \dots, T - 1$  do
4:    $x^{t+1} = x^t - \gamma g^t$ 
5:   for  $i \in \mathcal{G}$  do
6:      $g_i^{t+1} = \text{clip}_\tau(\nabla f_i(x^{t+1}, \xi_i^{t+1}))$ 
7:      $\omega_i^{t+1} \sim \mathcal{N}(0, \sigma_\omega^2 \mathbf{I})$ 
8:      $m_i^{t+1} = g_i^{t+1} + \omega_i^{t+1}$ 
9:   end for
10:  for  $i \in \mathcal{B}$  do
11:     $m_i^{t+1} = (*)$  sends arbitrary vector
12:  end for
13:   $g^{t+1} = \text{RAgg}(m_1^{t+1}, \dots, m_n^{t+1})$ 
14: end for

```

Algorithm 4 Safe-DSHB

```

1: Input:  $x^0 \in X$ , momentum  $\beta \in (0, 1]$ , step-size  $\gamma > 0$ ,
    $m_i^0 \in \mathbb{R}^d$ , clipping  $\tau > 0$ , DP noise variance  $\sigma_\omega^2 \geq 0$ 
2: for  $t = 0, \dots, T - 1$  do
3:    $x^{t+1} = x^t - \gamma g^t$ 
4:   for  $i \in \mathcal{G}$  do
5:      $g_i^{t+1} = \text{clip}_\tau(f_i(x^{t+1}, \xi_i^{t+1}))$ 
6:      $\omega_i^{t+1} \sim \mathcal{N}(0, \sigma_\omega^2 \mathbf{I})$ 
7:      $m_i^{t+1} = (1 - \beta)m_i^t + \beta(g_i^{t+1} + \omega_i^{t+1})$ 
8:   end for
9:   for  $i \in \mathcal{B}$  do
10:     $m_i^{t+1} = (*)$  sends arbitrary vector
11:   end for
12:    $g^{t+1} = \text{RAgg}(m_1^{t+1}, \dots, m_n^{t+1})$ 
13: end for

```

This finalizes the transition step of induction. The result of the theorem follows by setting $K = T - 1$. Indeed, from (70) we obtain

$$\frac{\gamma}{2} \sum_{t=0}^K \|\nabla f(x^t)\|^2 \leq \tilde{\Phi}^0 - \tilde{\Phi}^{K+1} + \tilde{\Delta} \leq 2\tilde{\Delta} \Rightarrow \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|^2 \leq \frac{4\tilde{\Delta}}{\gamma T}. \quad (115)$$

Final rate. We highlight that we are interested in the functional dependency of the rate on the problem constants. Therefore, in the rest of the proof, we omit using numerical constants. Translating momentum restriction to step-size restriction

$$\gamma = \frac{1}{L} \tilde{\mathcal{O}} \left(\min \left\{ 1, \underbrace{\left(\frac{L\tilde{\Delta}G}{T\sigma^2} \right)^{1/2}}_{\text{from term I (112)}}, \underbrace{\left(\frac{L\tilde{\Delta}\sqrt{G}}{\sigma\sqrt{T}(\sqrt{L\tilde{\Delta}} + \sigma/\sqrt{G} + \sqrt{c\delta_{\text{byz}}\zeta} + \sqrt{c\delta_{\text{byz}}\sigma})} \right)}_{\text{from term II (114)}}, \underbrace{\left(\frac{L\tilde{\Delta}}{c\delta_{\text{byz}}T\zeta^2} \right)}_{\text{from term III (113)}} \right\} \right), \quad (116)$$

We plug in the restrictions (116) on the step-size into (115) and obtain the rate

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|^2 &\leq \frac{L\tilde{\Delta}}{T} \tilde{\mathcal{O}} \left(1 + \left(\frac{T\sigma^2}{L\tilde{\Delta}G} \right)^{1/2} + \frac{\sigma\sqrt{T}(\sqrt{L\tilde{\Delta}} + \sigma/\sqrt{G} + \sqrt{c\delta_{\text{byz}}\zeta} + \sqrt{c\delta_{\text{byz}}\sigma})}{L\tilde{\Delta}\sqrt{G}} + \frac{c\delta_{\text{byz}}T\zeta^2}{L\tilde{\Delta}} \right) \\ &= \tilde{\mathcal{O}} \left(\frac{L\tilde{\Delta}}{T} + \left(\frac{\sigma^2 L\tilde{\Delta}}{GT} \right)^{1/2} + \frac{\sigma(\sqrt{L\tilde{\Delta}} + \sigma/\sqrt{G} + \sqrt{c\delta_{\text{byz}}\zeta} + \sqrt{c\delta_{\text{byz}}\sigma})}{\sqrt{GT}} + c\delta_{\text{byz}}\zeta^2 \right). \end{aligned} \quad (117)$$

It remains to provide a valid upper bound $\tilde{\Delta}$ on $\tilde{\Phi}_0$. Since we initialize $v_i^0 = 0$, then

$$\tilde{\Phi}^0 = F^0 + \frac{\gamma}{\beta} \|\nabla f(x^0)\|^2 \leq F^0 + \frac{1}{4L} \cdot 2L(f(x^0) - f^*) = 2F^0.$$

This concludes the proof. □

F ADDITIONAL EXPERIMENTS AND TRAINING DETAILS

Our implementation builds on the codebase of Horváth and Richtárik [2020] to simulate a distributed environment, while the attack implementations are adopted from Gorbunov et al. [2022b]. It can be found via the link <https://github.com/SanyaGaponov/ByzClip21SGD2M>.

Algorithm 5 Byz-Clip21-SGD2M+

```
1: Input:  $x^0 \in X$ , momentum parameters  $\beta, \hat{\beta} \in (0, 1]$ , step-size  $\gamma > 0$ ,  $g_i^0 = m_i^0 \in \mathbb{R}^d$ , clipping parameters  $\tau_i, \tau_o > 0$ ,  
   DP noise variance  $\sigma_\omega^2 \geq 0$   
2: for  $t = 0, \dots, T - 1$  do  
3:    $x^{t+1} = x^t - \gamma g^t$   
4:   for  $i \in \mathcal{G}$  do  
5:     Sample  $\mathcal{S}_i^{t+1} \sim \text{Unif}[m]$  of cardinality  $S$   
6:      $\omega_i^{t+1} \sim \mathcal{N}(0, \sigma_\omega^2 \mathbf{I})$   
7:      $v_i^{t+1} = (1 - \beta)v_i^t + \frac{\beta}{S} \sum_{j \in \mathcal{S}_i^{t+1}} (\text{clip}_{\tau_i}(\nabla f_{ij}(x^{t+1})) + \omega_i^{t+1})$   
8:      $c_i^{t+1} = \text{clip}_\tau(v_i^{t+1} - g_i^t)$   
9:      $g_i^{t+1} = g_i^t + \hat{\beta} \text{clip}_{\tau_o}(v_i^{t+1} - g_i^t)$   
10:  end for  
11:  for  $i \in \mathcal{B}$  do  
12:     $c_i^{t+1} = (*)$  sends arbitrary vector  
13:  end for  
14:   $m_i^{t+1} = m_i^t + \hat{\beta} c_i^{t+1}$   
15:   $g^{t+1} = \text{RAgg}(m_1^{t+1}, \dots, m_n^{t+1})$   
16: end for
```

F.1 TRAINING DETAILS

In our experiments, we compare Byz-Clip21-SGD2M against Byz-Clip-SGD and Safe-DSHB [Allouah et al., 2023b], outlined in Algorithm 3 and Algorithm 4, respectively. Note that Byz-Clip-SGD is a variant of DP-SGD in which a robust aggregation rule replaces server-side averaging. Since DP-SGD is state-of-the-art for private learning, this modification (Algorithm 3) serves as our first baseline. The method of Allouah et al. [2023b] originally performs per-example clipping (line 5); to align with our analysis under σ -sub-Gaussian noise, i.e., without assuming a finite-sum structure for f_i , we instead clip the stochastic gradient.

F.2 AMPLIFICATION BY SUB-SAMPLING

We next consider a modified variant of Byz-Clip21-SGD2M (see Algorithm 5), which we call Byz-Clip21-SGD2M+. Unlike the population setting considered in the main body, each client now holds a finite dataset of size m with a local objective

$$f(x) = \frac{1}{G} \sum_{i \in \mathcal{G}} f_i(x), \quad f_i(x) = \frac{1}{m} \sum_{j=1}^m f_{ij}(x).$$

At iteration t , client $i \in \mathcal{G}$ samples a mini-batch $\mathcal{S}_i^{t+1}$ of size S , averages example-wise clipped stochastic gradients, and updates its momentum buffer using a noised version of this average (line 7 in Algorithm 5). This modification enables privacy amplification through sub-sampling, as each per-example gradient is protected via the Gaussian mechanism, and local DP guarantees are implied by post-processing. Consequently, the required DP noise can be reduced to

$$\sigma_\omega = \frac{S}{m} \cdot \frac{\tau}{\varepsilon} \sqrt{T \log \frac{1}{\delta}},$$

following Abadi et al. [2016]. We highlight that Byz-Clip21-SGD2M+ has two clipping operators. In the experiments, we vary the inner one (line 7) while the outer one is fixed $\tau_o = 1$ (line 9). We compare this enhanced version of Byz-Clip21-SGD2M against Byz-Clip-SGD and original Safe-DSHB, where example-wise clipping is also incorporated into the algorithm design.

We compare all methods under the previous setup: CNN/MLP training on MNIST with a stronger privacy budget $\varepsilon \in \{0.1, 0.3, 1, 3\}$ and a label-flipping attack. We split the training set across $n = 25$ clients equally. We tune the learning rate over $\{10, 1, 0.1, 0.01\}$ and the clipping threshold over $\{1, 0.3, 0.1, 0.03, 0.01\}$ (this corresponds to τ_i in Byz-Clip21-SGD2M+). For Safe-DSHB and Byz-Clip21-SGD2M+ we set $\beta = 0.1$ as before; for Byz-Clip21-SGD2M+ we use $\tau_o = 1$ and $\hat{\beta} = 0.01$. To defend against Byzantine clients, we employ a hybrid aggregation rule on the server that combines NNM and CM. We decrease the number of epochs to 30.

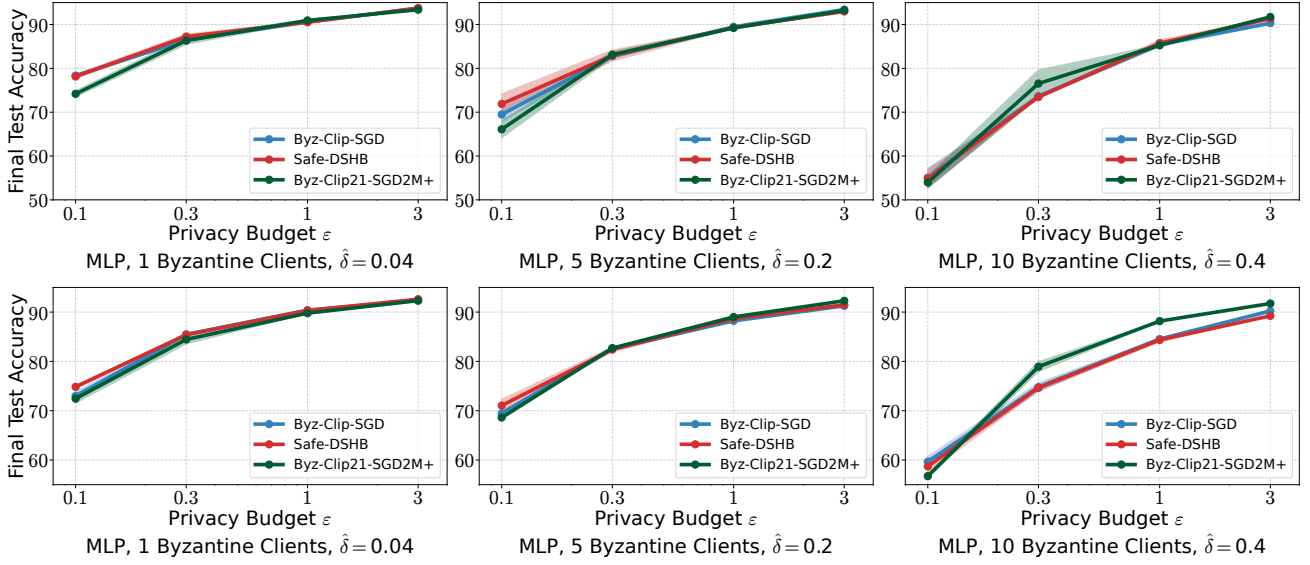


Figure F.1: Performance of Byz-Clip21-SGD2M+, Byz-Clip-SGD, and Safe-DSHB when training CNN (top line) and MLP (bottom line) models on the MNIST dataset for different numbers of Byzantine clients and privacy budgets, when Byzantine clients use a label flipping attack, when amplification by sub-sampling is done.

Results in Figure F.1 (bottom line) show that with 1 or 5 Byzantine clients, the methods are broadly competitive when training MLP model, with Safe-DSHB slightly leading at $\epsilon = 0.1$. With 10 Byzantines, Safe-DSHB and Byz-Clip-SGD edge out Byz-Clip21-SGD2M+ at $\epsilon = 0.1$, whereas for $\epsilon \in \{0.3, 1, 3\}$ Byz-Clip21-SGD2M+ clearly outperforms both baselines.

For the CNN model (Figure F.1, top row), the trend is similar: Byz-Clip-SGD and Safe-DSHB perform slightly better at $\epsilon = 0.1$, whereas Byz-Clip21-SGD2M+ is marginally more effective when the number of Byzantine clients is large.

Experiments on CIFAR10 [Krizhevsky, 2009] with the CNN model showed no significant difference between Byz-Clip21-SGD2M+ and the baselines; therefore, we do not report these results.

While we do not provide convergence guarantees for Byz-Clip21-SGD2M+, our experiments show that it delivers competitive performance against other baselines when amplification by sub-sampling is enabled. A theoretical analysis is deferred to future work since the example-wise clipping makes the analysis substantially more involved than for Byz-Clip21-SGD2M.