Reconciling Communication Compression and Byzantine-Robustness in Distributed Learning

Diksha Gupta[†]
University of Virginia

Antonio Honsell
Inria

Chuan Xu
UniCA, Inria, CNRS, i3S

Nirupam Gupta
University of Copenhagen

Giovanni Neglia
Inria, UniCA

Abstract

Distributed learning enables scalable model training over decentralized data, but remains hindered by Byzantine faults and high communication costs. While both challenges have been studied extensively in isolation, their interplay has received limited attention. Prior work has shown that naively combining communication compression with Byzantine-robust aggregation can severely weaken resilience to faulty nodes. The current theoretical state-of-the-art, Byz-DASHA-PAGE, leverages a momentum-based variance reduction scheme to counteract the negative effect of compression noise on Byzantine robustness. In this work, we introduce RoSDHB, a new algorithm that integrates classical Polyak momentum with a coordinated compression strategy. Theoretically, RoSDHB matches the convergence guarantees of Byz-DASHA-PAGE under the standard (G, B)-gradient dissimilarity model, while relying on milder assumptions and requiring less memory and communication per client. Empirically, RoSDHB demonstrates stronger robustness while achieving substantial communication savings.

[†]Work conducted during a post-doc at UniCA.

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

1 INTRODUCTION

Distributed learning enables scalable model training over decentralized data by leveraging parallel computation across multiple nodes (workers). In the standard setting, each worker computes local updates on its data and transmits them to a central server, which aggregates these updates to refine a global model. The updated model is then redistributed to the workers, and the process repeats iteratively. This paradigm accelerates training and preserves data locality, but it remains vulnerable to attacks and communication bottlenecks.

A particular challenge arises from malicious or faulty workers, commonly referred to as *Byzantine* workers in distributed computing literature [Lamport et al., 1982]. Such workers can arbitrarily corrupt their updates, severely degrading the convergence and accuracy of the global model. This has motivated a large body of work on Byzantine-robust distributed learning [Blanchard et al., 2017a, El Mhamdi et al., 2018, Yin et al., 2018, Li et al., 2021, Pillutla et al., 2022, Karimireddy et al., 2022, Allouah et al., 2023a], which develops robust aggregation rules to mitigate the impact of corrupted updates during server-side aggregation. These defenses are especially critical in open or partially trusted environments, such as *federated learning* [McMahan et al., 2017], where the non-i.i.d. nature of data across devices makes distinguishing Byzantine updates from legitimate but heterogeneous ones particularly challenging.

Another major line of work focuses on improving the communication efficiency of distributed learning. Transmitting full model parameters or gradients from every worker to the server at each round is often impractical under bandwidth constraints. To address this, communication compression schemes

such as gradient sparsification [Stich et al., 2018] and quantization [Alistarh et al., 2017] have been widely studied (e.g., [Chen et al., 2021, Hamer et al., 2020, Reisizadeh et al., 2020, Richtárik et al., 2021, Oh et al., 2024]). While these methods effectively reduce communication costs, they have mostly been analyzed in isolation from adversarial robustness. The interaction between communication compression and Byzantine resilience remains relatively underexplored, with only a few notable attempts [Bernstein et al., 2018, Sattler et al., 2019, Gorbunov et al., 2023, Rammal et al., 2024].

The current theoretical state-of-the-art (SOTA) approach for combining Byzantine-robustness with communication compression in distributed learning is Byz-DASHA-PAGE [Rammal et al., 2024]. This algorithm considers a robust variant of distributed stochastic gradient descent (SGD) with *unbiased compression*, and addresses the adverse effect of compression on Byzantine-robustness through *momentum variance reduction*—a technique originally proposed in [Cutkosky and Orabona, 2019] to achieve optimal convergence rates under noisy gradients. This mechanism is central to the significant improvements of Byz-DASHA-PAGE over earlier methods such as Byz-VR-MARINA [Gorbunov et al., 2023], but it comes at the cost of maintaining additional momentum states *before compression*, i.e., on the worker side, which leads to substantial per-worker memory usage. Moreover, the convergence of Byz-DASHA-PAGE relies on a special property of loss functions, namely bounded global Hessian variance [Szlendak et al., 2021] (see Appendix A).

We address the aforementioned limitations of Byz-DASHA-PAGE by proposing a new algorithm, named **Ro**bust **S**parsified **D**istributed **H**eavy-**B**all (**RoSDHB**). The design of RoSDHB is deliberately simpler than Byz-DASHA-PAGE. Instead of the more involved *momentum variance reduction* scheme, RoSDHB employs the classical *Polyak's momentum* (a.k.a. heavy-ball method). Crucially, RoSDHB applies heavy-ball momentum *after* compression, i.e., once the compressed gradients have been received by the server. This design shift moves the momentum operation to the server and halves the memory requirements at each worker, while still effectively mitigating the noise introduced by compression. A further innovation of RoSDHB is the use of a *global mask*, i.e., a shared sparsity pattern applied across all workers. This design reduces the uplink communication footprint, as workers transmit only the jointly selected coordinates and not the corresponding indices. While offering both memory and communication improvements, the convergence guarantee of RoSDHB is comparable to that of Byz-DASHA-PAGE in the gradient descent setting, without the additional assumption of bounded global Hessian variance.

In particular, we establish tight convergence guarantees for RoSDHB on full batch gradients of smooth nonconvex loss functions under the standard (G, B)-gradient dissimilarity [Karimireddy et al., 2020] and Lipschitz smoothness assumptions. Unlike prior momentum-based results for Byzantine-robust distributed learning [Karimireddy et al., 2021, Farhadkhani et al., 2022], our analysis addresses sparsification noise whose variance grows with the gradient norm of the average loss, rather than being uniformly bounded. To handle this more challenging regime, we develop a new Lyapunov function, which could be of independent interest to the distributed optimization community.

Although RoSDHB achieves convergence rates comparable to Byz-DASHA-PAGE —under weaker and more standard assumptions—our experimental results reveal a clear practical advantage, in both full gradient and stochastic gradient scenarios. In particular, RoSDHB demonstrates stronger robustness to Byzantine attackers and achieves significant communication savings across benchmark image classification tasks.

In summary, our contributions are three-fold: (i) We propose a new distributed learning algorithm RoSDHB that provides Byzantine-robustness while enforcing communication compression. Compared to prior work, RoSDHB incurs reduced memory and communication costs for the workers. (ii) We show (for the first time) that simple heavy-ball momentum, applied after sparsification and in combination with a global mask, yields tight Byzantine-robustness under standard assumptions. (iii) Our experiments show that RoSDHB consistently outperforms Byz-DASHA-PAGE in practice, achieving superior robustness to Byzantine attackers with significant communication savings.

Theoretical novelty Our proof is inspired by the convergence analysis in [Allouah et al., 2023a], but there are two key differences. First, unlike in [Allouah et al., 2023a], the variance of the noise induced by random sparsification is not uniformly bounded; instead, it grows with the squared norm of the average honest gradient (cf. Lemma B.1 and Eq. (28) in [Allouah et al., 2023a]). Following the analysis in [Allouah et al., 2023a] would therefore yield a suboptimal error rate. Second, to obtain a tighter convergence error rate—matching the theoretical SOTA Byz-DASHA-PAGE—we design a new Lyapunov function with an additional term that depends on the momentum drift (cf. the proof sketch of Theo-

rem 1 and Eq. (40) in [Allouah et al., 2023a]).

The rest of the paper is organized as follows. Section 2 reviews background material on Byzantine-robustness and gradient sparsification. Section 3 introduces our algorithm RoSDHB, presents its theoretical analysis, and includes a formal comparison with Byz-DASHA-PAGE. Section 4 reports empirical results on benchmark machine learning tasks. Section 5 discusses the possible extensions of RoSDHB. Finally, Section 6 summarizes our contributions, and outlines promising directions for future research.

2 PROBLEM SETTING & BACKGROUND

In this section, we present the problem setup and the assumptions used throughout the paper. We use $\|\cdot\|$ to denote the Euclidean norm.

Distributed learning. We consider a server-based distributed learning setting with one central server and n workers $[n] \coloneqq \{1, 2, \ldots, n\}$. Each worker i holds a local dataset $D_i = \{z_i^1, z_i^2, \ldots, z_i^m\}$, drawn from a possibly different distribution, as is often the case in *federated learning* settings. For each data point $z \in \mathcal{Z}$, a model with parameter vector $\theta \in \mathbb{R}^d$ incurs a loss $\ell(\theta, z)$, where ℓ is real-valued and differentiable. The local empirical loss at worker i is $\mathcal{L}_i(\theta) = \frac{1}{m}\sum_{j=1}^m \ell(\theta, z_i^j)$, and the global objective is to minimize the average loss $\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^n \mathcal{L}_i(\theta)$.[1] We allow $\mathcal{L}(\theta)$ to be nonconvex, in which case the goal is to find a stationary point (defined formally later). To this purpose, the server can employ distributed gradient descent (DGD). At each iteration $t \in 1, \ldots, T$, the server broadcasts the current model θ^{t-1} to all workers. Each worker i computes its local gradient $g_i^t = \nabla \mathcal{L}_i(\theta^{t-1})$ and sends it to the server. The server aggregates them as $g^t = \frac{1}{n}\sum_{i=1}^n g_i^t$ and updates the model by $\theta^t = \theta^{t-1} - \gamma g^t$, where $\gamma > 0$ is the learning rate. The initial model θ^0 is chosen arbitrarily.

Communication Compression. To reduce the communication cost of DGD, several compression schemes have been proposed [Beznosikov et al., 2023]. We focus on *RandK sparsification* [Stich et al., 2018], where at each iteration t worker i sends only k ($k \leq d$) randomly selected coordinates of its local gradient g_i^t. We define the *compression parameter* as $\alpha \coloneqq d/k$. Formally, each worker samples a binary mask $\mathsf{mask}_i(k) \in \{0,1\}^d$ with exactly k entries equal to 1, indicating the selected coordinates. The compressed message is the pair $(\mathsf{mask}_i(k), \mathcal{C}_k(g_i^t))$, where $\mathcal{C}_k(g_i^t) = ([g_i^t]_{\ell_1}, \ldots, [g_i^t]_{\ell_k})$ contains the gradient values at the selected indices $\ell_1 < \cdots < \ell_k$ satisfying $[\mathsf{mask}_i(k)]_{\ell_j} = 1$. Using this information, the server reconstructs $\tilde{g}_i^t = \alpha(g_i^t \odot \mathsf{mask}_i(k)) \in \mathbb{R}^d$, where $\odot$ denotes the element-wise (Hadamard) product. The server then averages $\overline{\tilde{g}}^t = \frac{1}{n}\sum_{i=1}^n \tilde{g}_i^t$ and updates the model as $\theta^t = \theta^{t-1} - \gamma \overline{\tilde{g}}^t$. RandK is an unbiased compression operator with bounded variance [Beznosikov et al., 2023, Def. 1], since

$$\mathbb{E}\left[\tilde{g}_i^t\right] = g_i^t \quad \text{and} \quad \mathbb{E}\left[\left\|\tilde{g}_i^t - g_i^t\right\|^2\right] \leq (\alpha - 1)\left\|g_i^t\right\|^2,$$

where the expectation is over the randomness of $\mathsf{mask}_i(k)$.

Byzantine-robustness. We consider a scenario wherein an adversary corrupts a set of at most f workers out of total n, whose identity is fixed but a priori unknown to the server. We refer to these corrupted workers as *Byzantine workers*, as per the terminology used in distributed computing [Lamport et al., 1982], and the remaining uncorrupted workers as the *honest workers*. We denote by $\mathcal{H} \subseteq [n]$ the set of honest workers ($|\mathcal{H}| = n - f$). Byzantine workers can deviate arbitrarily from the prescribed algorithm. We consider the worst-case scenario where they may collude seamlessly, know the learning algorithm used by the server, and observe all messages exchanged between the server and the honest workers throughout training. In the presence of Byzantine workers, the server's objective is to find a stationary point of the average loss over the honest workers, $\mathcal{L}_\mathcal{H}(\theta) \coloneqq \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \mathcal{L}_i(\theta)$. An algorithm that, in the presence of at most f Byzantine workers, guarantees convergence to an ϵ-stationary point of $\mathcal{L}_\mathcal{H}$ is said to be (f, ϵ)-resilient. Formally [Allouah et al., 2023a]:

Definition 2.1 ((f, ϵ)-Resilience)**.** *A distributed learning algorithm is said to be (f, ϵ)-resilient if, in the presence of at most f Byzantine workers, it outputs a parameter $\hat{\theta} \in \mathbb{R}^d$ such that $\left\|\nabla \mathcal{L}_\mathcal{H}(\hat{\theta})\right\|^2 \leq \epsilon$.*

The convergence of distributed first-order methods typically relies on standard conditions concerning the regularity of the global loss and the heterogeneity of local losses. In this paper, we adopt two broad assumptions that capture these aspects. These assumptions are imposed only on the honest workers, while Byzantine workers are left unconstrained and may act arbitrarily to disrupt training. We first introduce the following smoothness assumption on the honest workers' global loss, which is standard even in centralized first-order machine learning methods [Bottou et al., 2018].

Assumption 1. *(Smoothness) The average loss function $\mathcal{L}_\mathcal{H} : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz smooth, i.e., there*

[1] The analysis remains valid both when workers hold datasets of different sizes and when the global loss is defined as a weighted combination of the local losses.

exists $L \geq 0$ such that for all $\theta, \theta' \in \mathbb{R}^d$,

$$\|\nabla \mathcal{L}_{\mathcal{H}}(\theta) - \nabla \mathcal{L}_{\mathcal{H}}(\theta')\| \leq L \|\theta - \theta'\| .$$

To model data heterogeneity across honest workers, we rely on the standard notion of (G, B)-gradient dissimilarity [Karimireddy et al., 2020, Allouah et al., 2023b].

Assumption 2. *((G, B)-gradient dissimilarity) There exist two non-negative constants G and B, such that, for all $\theta \in \mathbb{R}^d$, we have:*

$$\frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \|\nabla \mathcal{L}_i(\theta) - \nabla \mathcal{L}_{\mathcal{H}}(\theta)\|^2 \leq G^2 + B^2 \|\nabla \mathcal{L}_{\mathcal{H}}(\theta)\|^2 .$$

DGD with simple averaging as the server-side aggregation rule, $g^t = \frac{1}{n} \sum_{i=1}^n g_i^t$, does not achieve (f, ϵ)-resilience [Blanchard et al., 2017b]. Resilience requires limiting the influence of Byzantine workers, which can be achieved, for example, by replacing averaging with a *robust aggregation* rule [Allouah et al., 2023a].

Definition 2.2 ((f, κ)-Robustness)**.** *An aggregation rule $F : (\mathbb{R}^d)^n \to \mathbb{R}^d$ is said to be (f, κ)-robust if, for any set of n vectors $\{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ and any subset $S \subseteq [n]$ of size $n - f$, it holds that*

$$\|F(x_1, \ldots, x_n) - \overline{x}_S\|^2 \leq \frac{\kappa}{|S|} \sum_{i \in S} \|x_i - \overline{x}_S\|^2 ,$$

where $\overline{x}_S \coloneqq \frac{1}{|S|} \sum_{i \in S} x_i$.

The above definition encompasses several robust aggregation rules, including coordinate-wise trimmed mean (CWTM), and geometric median (GeoMed) [Guerraoui et al., 2024, Chapter 4], which are (f, κ)-robust for different values of κ, where κ typically depends on f and n, and increases with f.

We call *robust DGD* a variant of DGD where averaging is replaced by an (f, κ)-robust aggregation rule. Specifically, at each iteration t, the server updates the parameter vector as $\theta^t = \theta^{t-1} - \gamma R^t$, where $R^t \coloneqq F(g_1^t, \ldots, g_n^t)$. Under Assumptions 1 and 2, it has been shown [Allouah et al., 2023b] that if $\kappa = \frac{f}{n-2f}$, robust DGD achieves (f, ϵ)-convergence with $\epsilon \in O\Big(\frac{f}{n-(2+B^2)f}\Big)$, which matches the lower bound on ϵ established in [Allouah et al., 2023a]. When $n \geq (2 + \nu)f$ for some $\nu > 0$, robust aggregation rules such as CWTM and GeoMed (when combined with nearest-neighbor mixing, NNM) achieve $\kappa \in \Theta\Big(\frac{f}{n-2f}\Big)$ [Allouah et al., 2023a].

Integrating Robustness and Communication Compression. Achieving both communication compression, to alleviate bandwidth constraints, and robustness, to mitigate the impact of Byzantine workers, is highly desirable but nontrivial. Simply adding compression schemes such as RandK to robust DGD is insufficient; more sophisticated algorithms are required. Some works attempt this integration but under restrictive assumptions: SignSGD with majority vote [Bernstein et al., 2018] requires honest gradients to follow identical unimodal distributions, while norm filtering with error feedback under biased compression [Ghosh et al., 2021] assumes subexponential gradients. The current state-of-the-art method addressing this joint goal is Byz-DASHA-PAGE [Rammal et al., 2024]. Its key ingredient is *momentum variance reduction*, a scheme originally proposed in [Cutkosky and Orabona, 2019] to attain optimal convergence rates for SGD on smooth nonconvex functions. Byz-DASHA-PAGE, however, comes with notable memory costs. Each worker must store four vectors of dimension d (the number of model parameters): the current parameter vector, the current gradient, the gradient from the previous communication round, and an additional pseudogradient. On the server side, the memory requirement is $2(n + 1)d$, as it must store each worker's pseudogradient and current update, along with the global model and the robust aggregate. Furthermore, under RandK sparsification, each worker uses local masks and therefore transmits k values together with the k corresponding indices (or, equivalently, d bits). Finally, the resilience guarantees of Byz-DASHA-PAGE rely not only on Assumptions 1 and 2, but also on the additional condition of *bounded global Hessian variance* [Szlendak et al., 2021] (reported as Assumption 3 in Appendix A). Unless each worker local loss is L-smooth, it is easy to construct an example when the global loss is smooth, but bounded global Hessian variance is not satisfied (see Lemma A.1 in Appendix A).

As we show in the following, our algorithm RoSDHB overcomes these limitations of Byz-DASHA-PAGE.

3 OUR ALGORITHM: ROBUST SPARSIFIED DISTRIBUTED HEAVY-BALL (ROSDHB)

In this section, we present our algorithm Robust Sparsified Distributed Heavy-Ball (RoSDHB). The pseudocode is provided in Algorithm 1, while Section 3.1 details its key components and complexity. Section 3.2 then presents the convergence analysis of RoSDHB and compares it with state-of-the-art results, including Byz-DASHA-PAGE.

3.1 Description of RoSDHB

In each iteration $t \in \{1, \ldots, T\}$, with $T \geq 1$ denoting the total number of iterations, the server broadcasts to

Algorithm 1: Robust **S**parsified **D**istributed **H**eavy-**B**all (**RoSDHB**)

Input: Initial model $\theta^0 \in \mathbb{R}^d$ (chosen arbitrarily), total iterations $T \geq 1$, learning rate γ, robust aggregator F, momentum coefficient $\beta \in [0,1)$, sparsification parameter k, and, for each honest worker w_i: $m_i^0 = 0$.

For $t = 1$ to T:

1: **Server** generates $\mathsf{mask}(k) \in \{0,1\}^d$ with k randomly chosen elements set to 1 and rest to 0.
2: **Server** broadcasts current model θ^{t-1} and $\mathsf{mask}(k)$ to all the workers.
3: For each **honest worker** i do:
   a. Compute local gradient: $g_i^t = \nabla \mathcal{L}_i\left(\theta^{t-1}\right)$.
   b. Send to the server: $\mathcal{C}_k\left(g_i^t\right) = \left([g_i^t]_{\ell_1}, \ldots, [g_i^t]_{\ell_k}\right)$ where $\ell_1 < \ldots < \ell_k$, $[\mathsf{mask}(k)]_{\ell_j} = 1, \forall j \in [k]$.
   // `A Byzantine worker` j `can send arbitrary` k `values in` $\mathcal{C}_k\left(g_j^t\right)$.
4: **Server** computes for each worker i, : $\tilde{g}_i^t = \frac{d}{k}(g_i^t \odot \mathsf{mask}(k))$ using $\mathcal{C}_k\left(g_i^t\right)$.
5: **Server** computes momentum for each worker i: $m_i^t = \beta m_i^{t-1} + (1-\beta)\tilde{g}_i^t$.
6: **Server** aggregates the momenta: $R^t = F(m_1^t, \ldots, m_n^t)$.
7: **Server** updates the model: $\theta^t = \theta^{t-1} - \gamma R^t$,

Server returns $\hat{\theta}$, chosen uniformly at random from $\{\theta^0, \ldots, \theta^{T-1}\}$.

all workers (i) the current model parameter θ^{t-1} and (ii) a RandK sparsification mask $\mathsf{mask}(k) \in \{0,1\}^d$, obtained by setting k randomly chosen entries to 1 and the rest to 0. The initial model θ^0 is chosen arbitrarily by the server. Upon receiving θ^{t-1} and the mask, each honest worker i computes its local gradient $g_i^t = \nabla \mathcal{L}_i(\theta^{t-1})$ and sends back the k coordinates corresponding to the nonzero entries of the mask, i.e., $\mathcal{C}_k(g_i^t) = ([g_i^t]_{\ell_1}, \ldots, [g_i^t]_{\ell_k})$ with $\ell_1 < \cdots < \ell_k$ and $[\mathsf{mask}(k)]_{\ell_j} = 1$ for all $j \in [k]$. In contrast, a Byzantine worker j may transmit any arbitrary set of k values in $\mathcal{C}_k(g_j^t)$. From the received messages, the server reconstructs an unbiased estimator of each local gradient:

$$\tilde{g}_i^t = \tfrac{d}{k}\left(g_i^t \odot \mathsf{mask}(k)\right).$$

The server then updates the momentum vector m_i^t of each worker as

$$m_i^t = \beta m_i^{t-1} + (1-\beta)\tilde{g}_i^t,$$

with momentum coefficient $\beta \in [0,1)$ and initialization $m_i^0 = 0$. Next, the server applies the robust aggregation rule F to obtain $R^t = F(m_1^t, \ldots, m_n^t)$, and updates the model as

$$\theta^t = \theta^{t-1} - \gamma R^t.$$

After T iterations, the server outputs $\hat{\theta}$, chosen uniformly at random from $\theta^0, \ldots, \theta^{T-1}$. The use of momentum vectors in the update rule mirrors Polyak's momentum method, also known as the heavy-ball method, which motivates the name of our algorithm.

RoSDHB complexity. RoSDHB is lightweight in terms of computation, memory, and communication. Each honest worker computes one local gradient per iteration, incurring the same computational cost as in standard DGD. On the memory side, each worker maintains only two d-dimensional vectors—the current parameter vector and gradient—in contrast to Byz-DASHA-PAGE, which requires four such vectors. The server stores the global model together with one momentum vector per worker, for a total memory footprint of $2(n+1)d$, identical to Byz-DASHA-PAGE. Communication from workers to the server is slightly reduced: with a global mask, each worker transmits only k gradient values per iteration without sending the corresponding indices, whereas Byz-DASHA-PAGE with RandK requires transmitting both k values and their indices (or equivalently d bits). This saving comes at the cost of distributing the mask from the server to the workers. However, this overhead of transmitting masks can be made negligible if they are generated by a random number generator shared between the server and the workers, initialized with a common seed, so that only the initial seed must be communicated. Overall, RoSDHB incurs strictly lower memory and worker-to-server communication overhead than Byz-DASHA-PAGE, while preserving the same computational complexity.

3.2 Analysis of RoSDHB

We prove the following theorem on the convergence of RoSDHB in Appendix B.

Theorem 1. *Under Assumptions 1 and 2, Algorithm 1 with an (f,κ)-robust aggregation rule F such that $\kappa B^2 \leq \frac{1}{7}$, a learning rate $\gamma \leq \frac{k}{dcL}$ (with $c = 23200$), and a momentum coefficient $\beta = \sqrt{1 - 24\gamma L}$, satisfies the following guarantee*

$$\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\hat{\theta})\right\|^2\right] \leq \frac{45\left(\mathcal{L}_{\mathcal{H}}(\theta^0) - \mathcal{L}_{\mathcal{H}}^*\right)}{\gamma T(1-\kappa B^2)} + \frac{216\kappa G^2}{1-\kappa B^2},$$

where $\mathcal{L}_{\mathcal{H}}^* = \min_{\theta \in \mathbb{R}^d} \mathcal{L}_{\mathcal{H}}(\theta)$.

Proof sketch. The proof relies on the following novel Lyapunov function:

$$V^t := \mathbb{E}\big[2\mathcal{L}_{\mathcal{H}}(\theta^{t-1}) + \tfrac{1}{8L}\|\delta^t\|^2 + \tfrac{\kappa}{4L}\Upsilon_{\mathcal{H}}^{t-1}\big]\,.$$

Here $\delta^t := \overline{m}^t - \nabla\mathcal{L}_{\mathcal{H}}(\theta^{t-1})$, with $\overline{m}^t := \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} m_i^t$, measures the deviation of the averaged momentum from the actual gradient. The term $\Upsilon_{\mathcal{H}}^{t-1} := \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\|m_i^{t-1} - \overline{m}_{\mathcal{H}}^{t-1}\|^2$ instead captures the drift of individual client momenta from their average. ■

Recall that $\alpha = d/k$ denotes the RandK compression ratio. We obtain the following corollary.

Corollary 1. *In the same conditions of Theorem 1, let* $\gamma = \frac{1}{\alpha c L}$*, then*

$$\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}(\hat{\theta})\right\|^2\right] \leq \mathcal{O}\Big(\tfrac{\alpha}{T(1-\kappa B^2)} + \tfrac{\kappa G^2}{1-\kappa B^2}\Big)\,. \qquad (1)$$

On the tightness of our result. When $\alpha = 1$ (i.e., no compression), Corollary 1 recovers the same bound as the SOTA result without compression [Allouah et al., 2023b]. Hence, the reasoning in Sec. 5.1 of [Allouah et al., 2023b] directly implies that the convergence bound in Corollary 1 is optimal under the best possible robust aggregation rule, namely one with $\kappa = f/(n-2f)$. This result also shows that the non-vanishing error term $\frac{\kappa G^2}{1-\kappa B^2}$ is unavoidable in Byzantine-robust settings unless additional assumptions are imposed, such as homogeneous data ($G = 0$) or the absence of adversaries ($f = 0$).

Interplay between compression and robustness. The corollary also highlights an interplay between compression and robustness. When $B = 0$, it separates the two effects in a manner analogous to that in [Allouah et al., 2023a] under $(G, 0)$-heterogeneity: one term reflects the impact of compression, and the other the impact of Byzantine robustness. By contrast, when $B > 0$, the first term in the bound shows that compression and robustness no longer decouple but instead amplify each other's effect: stronger compression (higher α) increases the influence of Byzantine robustness, and a larger fraction of Byzantine workers exacerbates the impact of compression.

Comparison with Byz-DASHA-PAGE. The SOTA method addressing both robustness and compression is Byz-DASHA-PAGE [Rammal et al., 2024]. Our algorithm's convergence guarantee in Theorem 1 matches that of Byz-DASHA-PAGE, but under fewer assumptions. In particular, our analysis requires only Lipschitz smoothness of the honest-average loss, whereas Byz-DASHA-PAGE further assumes *bounded global Hessian variance* for the honest workers' losses. For a fair comparison, we adapt the guarantee reported in [Rammal et al., 2024] to the gradient–descent setting (with $n - f \geq 2$) and evaluate it in the most favorable regime for Byz-DASHA-PAGE —namely, setting the Hessian-variance term to zero and omitting some terms. Let $\tilde{\theta}$ denote its output after T iterations. The resulting *best-case* upper bound, derived in Appendix C, is:

$$\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\tilde{\theta}\right)\right\|^2\right] \lesssim \mathcal{O}\left(\tfrac{1+4(\alpha-1)\left(\frac{1}{\sqrt{n}}+\sqrt{\kappa}\right)}{T(1-2\kappa B^2)} + \tfrac{\kappa G^2}{1-2\kappa B^2}\right).$$

Comparing the above rate with the bound for RoSDHB in Corollary 1, two differences stand out. First, Byz-DASHA-PAGE has a worse dependence on the robustness parameter κ: both terms are divided by $(1-2\kappa B^2)$ instead of $(1-\kappa B^2)$ as in RoSDHB, and the vanishing term carries an additional $\sqrt{\kappa}$ factor. As $\kappa \geq f/(n-2f)$, this penalty grows with the Byzantine fraction. Second, the impact of compression ratio α is milder for Byz-DASHA-PAGE when f/n is small, because α appears only through $(\alpha-1)(1/\sqrt{n}+\sqrt{\kappa})$, whereas for RoSDHB the $1/T$ term scales linearly with α. Consequently, since $\kappa \in \mathcal{O}(f/n)$ for optimal robust aggregation [Allouah et al., 2023a], the former incurs lighter compression penalty in small-f, large-n regime.

4 EMPIRICAL EVALUATION

4.1 Experimental setup

Datasets and Models. We evaluate our method on two standard image classification benchmarks: MNIST and CIFAR-10. The MNIST dataset [Deng, 2012] consists of grayscale images of handwritten digits, with 60,000 training and 10,000 test samples across 10 digit classes. CIFAR-10 [Krizhevsky et al., 2014] contains 32×32 RGB images, comprising 50,000 training and 10,000 test samples over 10 object categories. For MNIST, we use a lightweight neural network with one convolutional layer followed by a fully connected layer, totaling about 12K parameters. For CIFAR-10, we employ a deeper ResNet-18 model [He et al., 2016], with about 11M parameters.

Distributed system setup. We consider a server-based distributed learning scenario with 10 honest workers and f Byzantine attackers. To model real-world data heterogeneity, we distribute samples with the same label across clients following a symmetric Dirichlet distribution with parameter $w = 5$ [Wang et al., 2020]. The server applies the coordinate-wise trimmed mean (CWTM) robust aggregation method [Yin et al., 2018], while

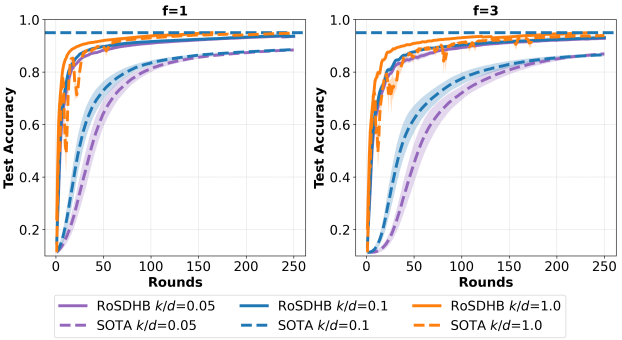
clients compress their updates using the RandK sparsifier [Beznosikov et al., 2023], using a sampling ratio $1/\alpha = k/d$ that retains only k out of d coordinates. We evaluate resilience against two attacks: the Fall of Empires (FOE) [Xie et al., 2020] and A Little Is Enough (ALIE) [Baruch et al., 2019]. Detailed descriptions of CWTM, FOE and ALIE are in Appendix D.1.

Hyperparameters. Learning rates are tuned via grid search for each compression ratio in the non-adversarial setting ($f = 0$), for both RoSDHB and Byz-DASHA-PAGE. Refer to Appendix C for the full description of Byz-DASHA-PAGE. In the MNIST experiments, each honest worker computes the full gradient, ensuring consistency with the theoretical analysis of RoSDHB. For the more memory-intensive training of ResNet-18 on CIFAR-10, we use mini-batch gradient updates, as is standard in practice. RoSDHB honest workers compute mini-batch gradient updates with $b = 128$ at each iteration. In its general form, Byz-DASHA-PAGE allows each client to perform a full-gradient update with probability p and a mini-batch update with probability $1 - p$. We set the mini-batch size to $b = 128$ and follow the configuration in [Rammal et al., 2024], by setting $p = b/m$, where m is the local dataset size. Appendix D.2 provides all hyper-parameters and computing resources.

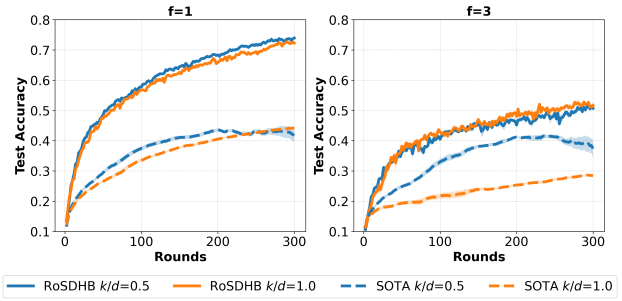
For MNIST, we evaluate scenarios with $f \in \{1, 3, 5\}$ Byzantine attackers and sampling ratios $1/\alpha = k/d \in \{0.05, 0.1, 0.3, 0.5, 1.0\}$ for 250 communication rounds. For CIFAR-10, we evaluate $f \in \{1, 3, 5\}$ and $k/d \in \{0.25, 0.5, 0.75, 1.0\}$ for 300 communication rounds. For each combination of f and k/d, we perform 5 independent runs and report the mean along with its standard error. Due to space limitations, we only report $f \in \{1, 3\}$ scenarios employing FOE attack strategy; the full set of experiments is provided in Appendix D.4. In the figures, Byz-DASHA-PAGE is simply indicated as SOTA. The code is available in `https://github.com/antoniohonsell/CC-and-BR-in-Distributed-Learning`.

4.2 Empirical results

Stable and fast convergence. Figure 1 illustrates the test accuracy convergence of our method RoSDHB (solid lines) compared with Byz-DASHA-PAGE (dashed lines) under different numbers of Byzantine attackers f and sampling ratios k/d for both MNIST and CIFAR-10 datasets. Across all experimental settings, RoSDHB exhibits faster and more stable convergence than Byz-DASHA-PAGE. For instance, for $k/d \leq 0.1$, RoSDHB achieves over 92.5% test accuracy on MNIST by the end of training, compared to only about 88.5% for the baseline. On CIFAR-10, RoSDHB attains over 70% test accuracy, whereas Byz-DASHA-PAGE reaches only around 44%. Moreover, the convergence curves of RoSDHB remain relatively close across different k/d values, indicating that its performance is less sensitive to the sampling ratio than the baseline.



(a) MNIST Dataset.



(b) CIFAR-10 Dataset.

Figure 1: The convergence plots of RoSDHB and Byz-DASHA-PAGE (SOTA) under varying number of Byzantine workers f and sampling ratios k/d.

Speedup and robustness. In Figure 2, we present the number of communication rounds required to reach different test accuracies under varying numbers of Byzantine attackers f. As expected, the number of communication rounds grows with f, reflecting the theoretical dependence of the convergence rate on κ, which itself increases with the number of Byzantine attackers f. Overall, RoSDHB achieves more than 4.5× speedup on MNIST with $k/d = 0.1$ and 2.3× on CIFAR-10 with $k/d = 0.5$, where speedup is measured as the reduction in communication rounds needed to reach a given test accuracy compared to Byz-DASHA-PAGE. The exact speedup values on different threshold values are reported in Tables 3 and 4 in Appendix D.3. Furthermore, RoSDHB shows stronger robustness to Byzantine attackers: with $f = 3$, it still outperforms the baseline with only $f = 1$. By contrast, the baseline with $f = 3$ fails to reach the target thresholds of 88% on MNIST and 44% on CIFAR-10 within the allotted communication rounds.

Compression & communication efficiency. Fig-

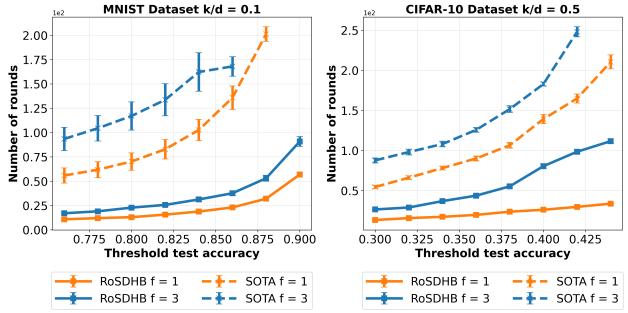
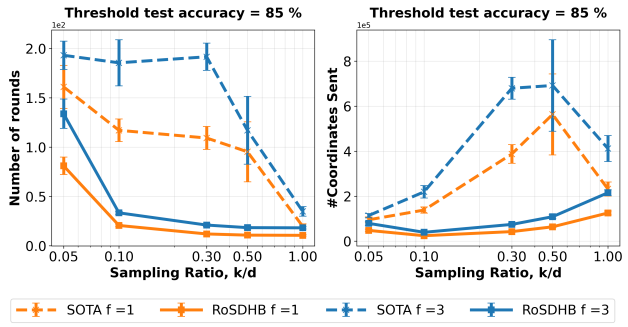
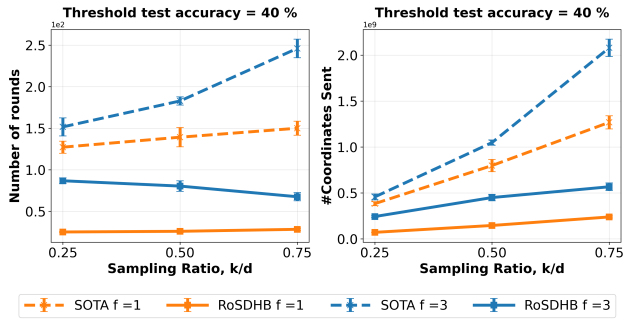


Figure 2: Comparison of RoSDHB and Byz-DASHA-PAGE (SOTA) convergence time required to reach various test threshold accuracy under varying Byzantine workers $f \in \{1, 3\}$.



(a) MNIST Dataset.



(b) CIFAR-10 Dataset.

Figure 3: Comparison of convergence time and communication cost between RoSDHB and Byz-DASHA-PAGE under varying sampling ratios and number of Byzantine workers $f \in \{1, 3\}$.

ure 3 reports the performance of RoSDHB and the baseline, measured by the number of rounds and the number of coordinates transmitted per client to reach a target test accuracy, under different sampling ratios k/d. For RoSDHB, the number of rounds increases as the sampling ratio k/d $(= 1/\alpha)$ decreases. The interplay between compression and robustness is also evident: at smaller sampling ratios, RoSDHB becomes more vulnerable to Byzantine attackers, leading to a comparatively larger increase in the number of rounds. Both effects are consistent with our theoretical analysis, in particular Corollary 1. Interestingly, although smaller sampling ratios require more rounds, the reduced number of coordinates transmitted per round can completely offset this increase (Fig. 3b, right), or lead to non-monotonic behavior (Fig. 3a, right). By contrast, the results for Byz-DASHA-PAGE are less aligned with its theoretical analysis in [Rammal et al., 2024]. While the expected trend holds for MNIST, with more rounds needed as the sampling ratio decreases, CIFAR-10 exhibits the opposite behavior. Furthermore, sensitivity to Byzantine attackers does not consistently increase at smaller sampling ratios. In any case, RoSDHB achieves substantial communication savings compared to the baseline across both datasets. For instance, RoSDHB reduces the number of transmitted coordinates by 89% relative to the baseline at $k/d = 0.3$ on MNIST. More detailed communication saving statistics are reported in Tables 5 and 6 in Appendix D.3.

Resilience to data heterogeneity. Finally, we examine the performance of RoSDHB and Byz-DASHA-PAGE under varying levels of data heterogeneity among honest workers. To this end, we control the Dirichlet distribution parameter w for data partitioning, ranging from a highly heterogeneous regime ($w = 0.5$) to a more moderate regime ($w = 5$). Figure 4 shows that RoSDHB consistently

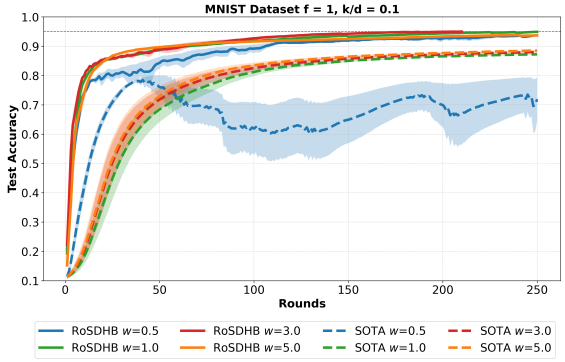


Figure 4: The convergence plot of RoSDHB and Byz-DASHA-PAGE (SOTA) on MNIST with $k/d = 0.1$ under varying data heterogeneity.

outperforms the baseline across all levels of data heterogeneity. Moreover, RoSDHB demonstrates greater resilience to heterogeneity: in the highly heterogeneous regime ($w = 0.5$), it achieves a test accuracy of 93.7%, whereas the baseline reaches only 71.9%.

The relative performance of RoSDHB and Byz-DASHA-PAGE is consistent under the ALIE attack (Appendix D.4). Overall, these results highlight the robustness and scalability of RoSDHB across a

wide range of compression ratios, Byzantine worker fractions, data heterogeneity levels, and benchmark datasets. Our proposed method not only sustains high accuracy under compression and adversarial conditions but also achieves significant improvements in communication efficiency.

5 DISCUSSION

5.1 Extension of bounds

Stochastic gradients Our theoretical bound is currently limited to full-gradient scenarios. Extending these results to stochastic gradients, however, presents several challenges. First, in the stochastic setting, RoSDHB is subject to multiple interacting noise sources, including stochastic gradient noise, compression noise from RandK, and Byzantine corruption controlled by the robust aggregator. While our Lyapunov function already tracks compression and robustness effects, incorporating the additional persistent martingale noise from SGD requires substantially more intricate variance tracking over time.
Second, momentum and variance accumulation introduce additional complexity. Because our momentum is applied post-compression and maintained per worker at the server, the stochastic setting would require controlling the accumulation of noise in the momentum variables (m_i^t), not just in the raw gradients. This is technically more subtle than in existing works where momentum is either absent or applied pre-compression.
Finally, a clean stochastic analysis would typically require additional assumptions, such as bounded variance or sub-Gaussian noise for each worker's stochastic gradient, on top of (G, B)-dissimilarity.

Unbiased compressor Although RoSDHB is described as using RandK, our theoretical results apply more broadly. In particular, the analysis remains valid (up to changes in some constants) for any unbiased sparsification operator C with bounded variance and satisfying the linearity property $C(\sum_i g_i^t) = \sum_i C(g_i^t)$. Beyond RandK, this includes, for example, schemes where the global mask is generated by sampling each coordinate j independently with probability p_j as in [Wangni et al., 2018].

5.2 Extension of RoSDHB to Biased Compressors

Extending RoSDHB to biased compressors via an error-feedback (EF) style correction, such as Byz-EF21 [Rammal et al., 2024], is an interesting direction for future research. However, this extension is not straightforward. For example, compressors such as Top-k require each client to select its own mask, whereas RoSDHB relies on a common global mask chosen by the server. Reconciling these two requirements would require new ideas and remains an open problem.

At the same time, using a global mask reduces the uplink communication footprint, since workers transmit only the jointly selected coordinates, without sending their indices. Besides, the global mask to provide additional robustness against Byzantine workers. Specifically, a global mask ensures that every active coordinate receives contributions from the same number of honest workers. This prevents situations—common with independently drawn local masks (e.g., Top-k)—where some coordinates receive very few honest updates. Such "weak spots" can be exploited by Byzantine workers, who may concentrate their manipulation on these underrepresented coordinates and thereby overpower coordinate-wise aggregation schemes such as CWTM.

We have performed preliminary experiments comparing RoSDHB with Byz-EF21 (see Appendix D.5). The results indicate that RoSDHB still outperforms Byz-EF21, despite using an unbiased compressor.

6 CONCLUDING REMARKS

We have proposed RoSDHB, a new distributed learning algorithm that achieves Byzantine-robustness while enforcing communication compression. By design, RoSDHB reduces both memory and communication costs at the workers compared to the SOTA Byz-DASHA-PAGE [Rammal et al., 2024]. Our theoretical analysis shows that the classical heavy-ball momentum, when applied after gradient sparsification and combined with a coordinated (global) sparsification scheme, is sufficient to guarantee tight Byzantine-robustness under standard assumptions. Beyond these guarantees, our experiments confirm that RoSDHB consistently outperforms Byz-DASHA-PAGE in practice, providing stronger resilience to Byzantine attackers together with substantial communication savings.

A natural future direction is to extend our framework to distributed stochastic gradient descent, moving beyond the full-gradient setting considered here. Other promising extensions include asynchronous training, adaptive sparsification masks, and more general compression schemes. Another open line of work is to explore the interplay of RoSDHB with privacy-preserving mechanisms such as differentially private Gaussian noise. Finally, motivated by federated learning, we plan to investigate the behavior of RoSDHB under multiple local SGD steps at each round.

ACKNOWLEDGMENTS

This work has been supported by the French government, through the France 2030 investment plan managed by the Agence Nationale de la Recherche, as part of the Université Côte d'Azur's Initiative of Excellence, reference ANR-15-IDEX-0001. This research was supported in part by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002 and in part by the European Network of Excellence dAIEDGE under Grant Agreement Nr. 101120726. This research was also supported in part by the EU HORIZON MSCA 2023 DN project FINALITY (G.A. 101168816)

References

[Alistarh et al., 2017] Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. (2017). Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in neural information processing systems*, 30.

[Allouah et al., 2023a] Allouah, Y., Farhadkhani, S., Guerraoui, R., Gupta, N., Pinot, R., and Stephan, J. (2023a). Fixing by mixing: A recipe for optimal Byzantine ml under heterogeneity. In Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 1232–1300. PMLR.

[Allouah et al., 2023b] Allouah, Y., Guerraoui, R., Gupta, N., Pinot, R., and Rizk, G. (2023b). Robust distributed learning: Tight error bounds and breakdown point under data heterogeneity. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[Baruch et al., 2019] Baruch, G., Baruch, M., and Goldberg, Y. (2019). A little is enough: Circumventing defenses for distributed learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Bernstein et al., 2018] Bernstein, J., Wang, Y.-X., Azizzadenesheli, K., and Anandkumar, A. (2018). signsgd: Compressed optimisation for non-convex problems. In *ICML*.

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

[Blanchard et al., 2017a] Blanchard, P., El Mhamdi, E. M., Guerraoui, R., and Stainer, J. (2017a). Machine learning with adversaries: Byzantine tolerant gradient descent. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 118–128, Red Hook, NY, USA. Curran Associates Inc.

[Blanchard et al., 2017b] Blanchard, P., El Mhamdi, E. M., Guerraoui, R., and Stainer, J. (2017b). Machine learning with adversaries: Byzantine tolerant gradient descent. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 119–129. Curran Associates, Inc.

[Bottou et al., 2018] Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311.

[Chen et al., 2021] Chen, M., Shlezinger, N., Poor, H. V., Eldar, Y. C., and Cui, S. (2021). Communication-efficient federated learning. *Proceedings of the National Academy of Sciences*, 118(17):e2024789118.

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

[Deng, 2012] Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142.

[El Mhamdi et al., 2018] El Mhamdi, E. M., Guerraoui, R., and Rouault, S. (2018). The hidden vulnerability of distributed learning in Byzantium. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3521–3530. PMLR.

[Farhadkhani et al., 2022] Farhadkhani, S., Guerraoui, R., Gupta, N., Pinot, R., and Stephan, J. (2022). Byzantine machine learning made easy by resilient averaging of momentums. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 6246–6283. PMLR.

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

[Gorbunov et al., 2023] Gorbunov, E., Horváth, S., Richtárik, P., and Gidel, G. (2023). Variance reduction is an antidote to Byzantines: Better rates, weaker assumptions and communication compression as a cherry on the top. In *The Eleventh International Conference on Learning Representations*.

[Guerraoui et al., 2024] Guerraoui, R., Gupta, N., and Pinot, R. (2024). *Robust Machine Learning*. Springer.

[Hamer et al., 2020] Hamer, J., Mohri, M., and Suresh, A. T. (2020). Fedboost: A communication-efficient algorithm for federated learning. In *International Conference on Machine Learning*, pages 3973–3983. PMLR.

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

[Karimireddy et al., 2021] Karimireddy, S. P., He, L., and Jaggi, M. (2021). Learning from history for Byzantine robust optimization. *International Conference On Machine Learning, Vol 139*, 139.

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

[Karimireddy et al., 2020] Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. (2020). Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR.

[Krizhevsky et al., 2014] Krizhevsky, A., Nair, V., and Hinton, G. (2014). The CIFAR-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 55(5).

[Lamport et al., 1982] Lamport, L., Shostak, R., and Pease, M. (1982). The Byzantine generals problem. *ACM Trans. Program. Lang. Syst.*, 4(3):382–401.

[Li et al., 2021] Li, T., Hu, S., Beirami, A., and Smith, V. (2021). Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pages 6357–6368. PMLR.

[McMahan et al., 2017] McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. In Singh, A. and Zhu, J., editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR.

[Nesterov et al., 2018] Nesterov, Y. et al. (2018). *Lectures on convex optimization*, volume 137. Springer.

[Oh et al., 2024] Oh, J., Lee, D., Won, D., Noh, W., and Cho, S. (2024). Communication-efficient federated learning over-the-air with sparse one-bit quantization. *IEEE Transactions on Wireless Communications*.

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

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

[Reisizadeh et al., 2020] Reisizadeh, A., Mokhtari, A., Hassani, H., Jadbabaie, A., and Pedarsani, R. (2020). Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *International conference on artificial intelligence and statistics*, pages 2021–2031. PMLR.

[Richtárik et al., 2021] Richtárik, P., Seidman, Y., Tang, R., Wang, X., Xu, C., and You, C. (2021). Ef21: A new, simpler, theoretically better, and practically faster error feedback. *NeurIPS*.

[Sattler et al., 2019] Sattler, F., Wiedemann, S., Müller, K.-R., and Samek, W. (2019). Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 31(9):3400–3413.

[Stich et al., 2018] Stich, S. U., Cordonnier, J.-B., and Jaggi, M. (2018). Sparsified sgd with memory. *Advances in neural information processing systems*, 31.

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

[Wang et al., 2020] Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D., and Khazaeni, Y. (2020). Federated learning with matched averaging. *ICLR*.

[Wangni et al., 2018] Wangni, J., Wang, J., Liu, J., and Zhang, T. (2018). Gradient sparsification for communication-efficient distributed optimization. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1306–1316.

[Xie et al., 2020] Xie, C., Koyejo, O., and Gupta, I. (2020). Fall of empires: Breaking Byzantine-tolerant sgd by inner product manipulation. In Adams, R. P. and Gogate, V., editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 261–270. PMLR.

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

Checklist

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

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

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

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

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

Reconciling Communication Compression and Byzantine-Robustness in Distributed Learning: Supplementary Materials

A About the Global Hessian-Variance Assumption

Assumption 3. *(Global Hessian-Variance). There exists a constant $L_\pm \geq 0$ such that for all $x, y \in \mathbb{R}^d$,*

$$\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\left\|\nabla\mathcal{L}_i(x)-\nabla\mathcal{L}_i(y)\right\|^2 \;-\; \left\|\nabla\mathcal{L}(x)-\nabla\mathcal{L}(y)\right\|^2 \;\leq\; L_\pm^2\left\|x-y\right\|^2. \tag{2}$$

In what follows, we provide a one-dimensional, two-worker construction in which the global objective $\mathcal{L}_H$ is L-smooth and satisfies Assumption 2, yet (2) fails for any finite $L_\pm$.

Lemma A.1 (Global L-smoothness without global Hessian-variance)**.** *Define*

$$\mathcal{L}_1(x) = \tfrac{1}{2}x^2 + \tfrac{1}{2}\sin x^2, \qquad \mathcal{L}_2(x) = \tfrac{1}{2}x^2 - \tfrac{1}{2}\sin x^2.$$

The average objective $\mathcal{L} = \frac{1}{2}(\mathcal{L}_1+\mathcal{L}_2)$ is globally 1-smooth and satisfies Assumption 2, but the functions do not satisfy Assumption 3.

Proof. Since $\mathcal{L}''(x) = \frac{1}{2}(\mathcal{L}_1''(x)+\mathcal{L}_2''(x)) = \frac{1}{2}\big((1+x\sin x)+(1-x\sin x)\big) = 1$, the function $\mathcal{L}$ is globally 1-smooth.

Note that $(\mathcal{L}_1'(x)-\mathcal{L}'(x))^2 = (\mathcal{L}_2'(x)-\mathcal{L}'(x))^2 = (x\cos x)^2 \leq x^2 = (\mathcal{L}'(x))^2$. Then Assumption 2 is satisfied with $B=0$ and $G=1$.

Let $\Delta_i(x,y) := \mathcal{L}_i'(x) - \mathcal{L}_i'(y)$. For $G=2$, the left hand side in (2) simplifies to

$$\frac{1}{2}\big(\Delta_1^2+\Delta_2^2\big) - \Big(\tfrac{\Delta_1+\Delta_2}{2}\Big)^2 = \frac{1}{4}\big(\Delta_1-\Delta_2\big)^2.$$

Moreover,

$$\Delta_1-\Delta_2 = 4\big(x\cos x - y\cos y\big).$$

If Assumption 3 held, then one would have, for some value $L_\pm$,

$$|x\cos x - y\cos y| \;\leq\; L_\pm\,|x-y|,$$

i.e., the function $g(x) = x\cos x$ should be $L_\pm$-Lipschitz. As the derivative of $g(\cdot)$ is unbounded ($g'(x) = -x\sin x+\cos x$), (2) cannot be satisfied for any $L_\pm$. ∎

B Proofs for RoSDHB

Notation. Let $\mathbb{E}_k[\cdot]$ denote the expectation over the random sparsification. Recall from Step (1) of Algorithm 1 that the RandK sparsifier $\mathsf{mask}(k) \in \{0,1\}^d$ is generated by the server and sent along with the model at the beginning of each iteration. This is used by all the workers in Step 3(b) for compressing the locally computed gradients. Thus, the expectation is over the independent random sampling of k coordinates, out of d coordinates, set by the server in a given iteration.

We first obtain some lemmas, which help bound the impact of global sparsification later in our convergence analysis.

Lemma B.1. *Consider Algorithm 1. For the set of honest workers $\mathcal{H}$, let $\overline{\tilde{g}}_{\mathcal{H}}^t = \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \tilde{g}_i^t$, denote the mean of the sparsified gradients of the honest workers at time t. Then, for any $t \in [T]$, we get*

$$\mathbb{E}_k\left[\left\|\overline{\tilde{g}}_{\mathcal{H}}^t - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right] \leq \left(\tfrac{d}{k}-1\right)\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2.$$

Proof. Recall from Algorithm 1 that all workers use the same set of sparsified bits in an iteration for compression, which were determined by the server. Fix a time $t \in [T]$, and let $\mathsf{mask}^t(k)$ be the mask set by the server for iteration t. Then,

$$\overline{\tilde{g}}_{\mathcal{H}}^t = \tfrac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \tilde{g}_i^t = \tfrac{d}{k}\tfrac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} g_i^t \odot \mathsf{mask}^t(k) = \tfrac{d}{k}\left(\tfrac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\nabla\mathcal{L}_i\left(\theta^{t-1}\right)\right)\odot\mathsf{mask}^t(k) = \tfrac{d}{k}\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\odot\mathsf{mask}^t(k).$$

Using the above, we have

$$\left\|\overline{\tilde{g}}_{\mathcal{H}}^t - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2 = \left\|\tfrac{d}{k}\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\odot\mathsf{mask}^t(k) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2.$$

Taking expectation over the random sparsification, $\mathbb{E}_k[\cdot]$ on both sides above, we get

$$\mathbb{E}_k\left[\left\|\overline{\tilde{g}}_{\mathcal{H}}^t - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right] \leq \mathbb{E}_k\left[\left\|\tfrac{d}{k}\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\odot\mathsf{mask}^t(k) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right].$$

Next, using the unbiased-ness of the random sparsifier, we get

$$\begin{aligned}
\mathbb{E}_k\left[\left\|\overline{\tilde{g}}_{\mathcal{H}}^t - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right] &\leq \tfrac{k}{d}\left\|\tfrac{d}{k}\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2 + \left(1-\tfrac{k}{d}\right)\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2 \\
&\leq \tfrac{k}{d}\left(\tfrac{d}{k}-1\right)^2\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2 + \left(1-\tfrac{k}{d}\right)\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2 \\
&= \left(\left(1-\tfrac{k}{d}\right)\left(\tfrac{d}{k}-1\right) + \left(1-\tfrac{k}{d}\right)\right)\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2 \\
&= \left(\tfrac{d}{k}-1\right)\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2,
\end{aligned}$$

which is the desired inequality. ∎

Next, we obtain some useful lemmas similar to previous work [Farhadkhani et al., 2022]. Let $m_{\mathcal{H}}^t = \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} m_i^t$.

Lemma B.2. *Consider Algorithm 1. Let $\mathcal{M}_i^t \coloneqq m_i^t - m_{\mathcal{H}}^t$ and $\Upsilon_{\mathcal{H}}^t \coloneqq \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\left\|\mathcal{M}_i^t\right\|^2$. Then, for any $t \in [T]$, we get*

$$\Upsilon_{\mathcal{H}}^t \leq \beta\Upsilon_{\mathcal{H}}^{t-1} + \left((1-\beta)^2\tfrac{d}{k} + \beta(1-\beta)\right)\left(G^2 + B^2\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right).$$

Proof. Consider an arbitrary $t \in [T]$ and an arbitrary $i \in \mathcal{H}$. Recall that $m_{\mathcal{H}}^t = \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} m_i^t$ and $m_i^t = \beta m_i^{t-1} + (1-\beta)\tilde{g}_i^t$, thus

$$m_{\mathcal{H}}^t = \tfrac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} m_i^t = \tfrac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\left(\beta m_i^{t-1} + (1-\beta)\tilde{g}_i^t\right) = \beta\overline{m}_{\mathcal{H}}^{t-1} + (1-\beta)\overline{\tilde{g}}_{\mathcal{H}}^t \tag{3}$$

Let $\mathcal{M}_i^t = m_i^t - \overline{m}_{\mathcal{H}}^t$. Then, substituting value of $m_{\mathcal{H}}^t$ from above and $m_i^t = \beta m_i^{t-1} + (1-\beta)\tilde{g}_i^t$,

$$\mathcal{M}_i^t = \left(\beta m_i^{t-1} + (1-\beta)\tilde{g}_i^t\right) - \left(\beta \overline{m}_{\mathcal{H}}^{t-1} - (1-\beta)\overline{\tilde{g}}_{\mathcal{H}}^t\right) = \beta \mathcal{M}_i^{t-1} + (1-\beta)(\tilde{g}_i^t - \overline{\tilde{g}}_{\mathcal{H}}^t) \tag{4}$$

Therefore,

$$\begin{aligned}
\left\|\mathcal{M}_i^t\right\|^2 &= \left\|\beta \mathcal{M}_i^{t-1} + (1-\beta)(\tilde{g}_i^t - \overline{\tilde{g}}_{\mathcal{H}}^t)\right\|^2 \\
&\leq \left\|\beta \mathcal{M}_i^{t-1}\right\|^2 + \left\|(1-\beta)(\tilde{g}_i^t - \overline{\tilde{g}}_{\mathcal{H}}^t)\right\|^2 + 2\beta(1-\beta)\left\langle \mathcal{M}_i^{t-1},\, (\tilde{g}_i^t - \overline{\tilde{g}}_{\mathcal{H}}^t)\right\rangle \\
&\leq \beta^2 \left\|\mathcal{M}_i^{t-1}\right\|^2 + (1-\beta)^2 \left\|(\tilde{g}_i^t - \overline{\tilde{g}}_{\mathcal{H}}^t)\right\|^2 + 2\beta(1-\beta)\left\langle \mathcal{M}_i^{t-1},\, (\tilde{g}_i^t - \overline{\tilde{g}}_{\mathcal{H}}^t)\right\rangle
\end{aligned}$$

Taking conditional expectation over the k-sparsifier, then using linearity of expectation and deterministic nature of m_i^t, we get:

$$\mathbb{E}_k\left[\left\|\mathcal{M}_i^t\right\|^2\right] \leq \beta^2 \left\|\mathcal{M}_i^{t-1}\right\|^2 + (1-\beta)^2 \mathbb{E}_k\left[\left\|\tilde{g}_i^t - \overline{\tilde{g}}_{\mathcal{H}}^t\right\|^2\right] + 2\beta(1-\beta)\left\langle \mathcal{M}_i^{t-1},\, \mathbb{E}_k\left[\tilde{g}_i^t - \overline{\tilde{g}}_{\mathcal{H}}^t\right]\right\rangle \tag{5}$$

Note that global sparsification uses the same random mask for all workers, thus due to unbiasedness of the random sparsifier, we have

$$\mathbb{E}_k\left[\tilde{g}_i^t - \overline{\tilde{g}}_{\mathcal{H}}^t\right] = g_i^t - \overline{g}_{\mathcal{H}}^t \quad \text{and} \quad \mathbb{E}_k\left[\left\|\tilde{g}_i^t - \overline{\tilde{g}}_{\mathcal{H}}^t\right\|^2\right] \leq \tfrac{d}{k}\left\|g_i^t - \overline{g}_{\mathcal{H}}^t\right\|^2.$$

Recall that $g_i^t = \nabla \mathcal{L}_i\left(\theta^{t-1}\right)$ and $\overline{g}_{\mathcal{H}}^t = \nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)$. Using this above, we get

$$\begin{aligned}
\mathbb{E}_k\left[\tilde{g}_i^t - \overline{\tilde{g}}_{\mathcal{H}}^t\right] &= \nabla \mathcal{L}_i\left(\theta^{t-1}\right) - \nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right), \\
\mathbb{E}_k\left[\left\|\tilde{g}_i^t - \overline{\tilde{g}}_{\mathcal{H}}^t\right\|^2\right] &\leq \tfrac{d}{k}\left\|\nabla \mathcal{L}_i\left(\theta^{t-1}\right) - \nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2.
\end{aligned} \tag{6}$$

Substituting from above in (5),

$$\begin{aligned}
\left\|\mathcal{M}_i^t\right\|^2 \leq \beta^2 \left\|\mathcal{M}_i^{t-1}\right\|^2 &+ (1-\beta)^2 \tfrac{d}{k}\left\|\nabla \mathcal{L}_i\left(\theta^{t-1}\right) - \nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2 \\
&+ 2\beta(1-\beta)\left\langle \mathcal{M}_i^{t-1},\, \nabla \mathcal{L}_i\left(\theta^{t-1}\right) - \nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle.
\end{aligned}$$

Due to Cauchy-Schwartz inequality, $2\langle a,\, b\rangle \leq 2\|a\|\|b\| \leq \|a\|^2 + \|b\|^2$. Thus, from above we obtain that

$$\begin{aligned}
\left\|\mathcal{M}_i^t\right\|^2 \leq \beta^2 \left\|\mathcal{M}_i^{t-1}\right\|^2 &+ (1-\beta)^2 \tfrac{d}{k}\left\|\nabla \mathcal{L}_i\left(\theta^{t-1}\right) - \nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2 \\
&+ \beta(1-\beta)\left(\left\|\mathcal{M}_i^{t-1}\right\|^2 + \left\|\nabla \mathcal{L}_i\left(\theta^{t-1}\right) - \nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right).
\end{aligned}$$

Upon re-arranging the right-hand side, we obtain that

$$\left\|\mathcal{M}_i^t\right\|^2 \leq \beta\left\|\mathcal{M}_i^{t-1}\right\|^2 + \left((1-\beta)^2 \tfrac{d}{k} + \beta(1-\beta)\right)\left\|\nabla \mathcal{L}_i\left(\theta^{t-1}\right) - \nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2.$$

Since the above holds for all $i \in \mathcal{H}$, taking the average on both sides over all $i \in \mathcal{H}$ yields

$$\tfrac{1}{\mathcal{H}}\textstyle\sum_{i\in\mathcal{H}}\left\|\mathcal{M}_i^t\right\|^2 \leq \beta\left(\tfrac{1}{\mathcal{H}}\sum_{i\in\mathcal{H}}\left\|\mathcal{M}_i^{t-1}\right\|^2\right) + \left((1-\beta)^2 \tfrac{d}{k} + \beta(1-\beta)\right)\left(\tfrac{1}{\mathcal{H}}\sum_{i\in\mathcal{H}}\left\|\nabla \mathcal{L}_i\left(\theta^{t-1}\right) - \nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right).$$

Substituting $\Upsilon_{\mathcal{H}}^t = \frac{1}{\mathcal{H}}\sum_{i\in\mathcal{H}}\left\|\mathcal{M}_i^t\right\|^2$ above, we obtain that

$$\Upsilon_{\mathcal{H}}^t \leq \beta\Upsilon_{\mathcal{H}}^{t-1} + \left((1-\beta)^2 \tfrac{d}{k} + \beta(1-\beta)\right)\left(\tfrac{1}{\mathcal{H}}\sum_{i\in\mathcal{H}}\left\|\nabla \mathcal{L}_i\left(\theta^{t-1}\right) - \nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right).$$

Finally, under the (G, B)-gradient dissimilarity property from Definition 2, we have

$$\Upsilon_{\mathcal{H}}^t \leq \beta\Upsilon_{\mathcal{H}}^{t-1} + \left((1-\beta)^2 \tfrac{d}{k} + \beta(1-\beta)\right)\left(G^2 + B^2\left\|\nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right),$$

which proves the lemma. ■

Lemma B.3. *Consider Algorithm 1, with F being (f,κ)-robust. Recall that $R^t \coloneqq F(m_1^t, ..., m_n^t)$. Denote $\mathcal{M}_i^t \coloneqq m_i^t - m_{\mathcal{H}}^t$ $\Upsilon_{\mathcal{H}}^t \coloneqq \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \left\|\mathcal{M}_i^t\right\|^2$, and $\xi^t \coloneqq R^t - m_{\mathcal{H}}^t$. Then, for each $t \in [T]$, we have:*

$$\left\|\xi^t\right\|^2 \leq \kappa\left(\beta\Upsilon_{\mathcal{H}}^{t-1} + \left((1-\beta)^2\,\tfrac{d}{k} + \beta(1-\beta)\right)\left(G^2 + B^2\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right)\right).$$

Proof. Fix an arbitrary $t \in [T]$. Since F is an (f,κ)-robust aggregator, then by Definition 2.2, we have:

$$\left\|R^t - m_{\mathcal{H}}^t\right\|^2 \leq \tfrac{\kappa}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\left\|m_i^t - m_{\mathcal{H}}^t\right\|^2 = \tfrac{\kappa}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\left\|\mathcal{M}_i^t\right\|^2 = \kappa\Upsilon_{\mathcal{H}}^t.$$

Using Lemma B.2 above, we get the desired inequality. ■

Next, we bound the momentum deviation of the honest workers from the mean gradient of loss of the honest workers for the case of global sparsification under the (G,B)-gradient dissimilarity condition.

Lemma B.4. *Suppose Assumption 1 holds true. Denote $\delta^t \coloneqq m_{\mathcal{H}}^t - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)$. Consider Algorithm 1 with $T > 1$, then for all $t \geq 2$, we have:*

$$\begin{aligned}\mathbb{E}\left[\left\|\delta^t\right\|^2\right] \leq{}& \zeta\,\mathbb{E}\left[\left\|\delta^{t-1}\right\|^2\right] + 2\gamma L\left(2\gamma L(1-\beta)^2\left(\tfrac{d}{k}-1\right) + (1+\gamma L)\beta^2\right)\mathbb{E}\left[\left\|\xi^{t-1}\right\|^2\right]\\ &+ \left(2(4\gamma^2L^2+1)(1-\beta)^2\left(\tfrac{d}{k}-1\right) + 4\gamma L(1+\gamma L)\beta^2\right)\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right)\right\|^2\right],\end{aligned}$$

where $\zeta = \left(8\gamma^2L^2(1-\beta)^2\left(\tfrac{d}{k}-1\right) + (1+4\gamma L)(1+\gamma L)\beta^2\right)$.

Proof. Fix an arbitrary $t \in [T]$. Then, substituting $m_{\mathcal{H}}^t = \beta\overline{m}_{\mathcal{H}}^{t-1} + (1-\beta)\overline{\tilde{g}}_{\mathcal{H}}^t$ in definition of δ^t:

$$\delta^t = m_{\mathcal{H}}^t - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right) = \beta\overline{m}_{\mathcal{H}}^{t-1} + (1-\beta)\overline{\tilde{g}}_{\mathcal{H}}^t - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)$$

Upon adding and subtracting $\beta\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)$ and $\beta\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right)$ on the R.H.S. above, we get:

$$\begin{aligned}\delta^t &= \beta\overline{m}_{\mathcal{H}}^{t-1} - \beta\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right) + (1-\beta)\overline{\tilde{g}}_{\mathcal{H}}^t - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right) + \beta\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right) + \beta\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right) - \beta\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\\ &= \beta\left(\overline{m}_{\mathcal{H}}^{t-1} - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right)\right) + (1-\beta)\left(\overline{\tilde{g}}_{\mathcal{H}}^t - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right) + \beta\left(\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right)\end{aligned}$$

Substituting $\delta^{t-1} = \overline{m}_{\mathcal{H}}^{t-1} - \beta\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right)$ above:

$$\delta^t = \beta\delta^{t-1} + (1-\beta)\left(\overline{\tilde{g}}_{\mathcal{H}}^t - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right) + \beta\left(\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right)$$

Taking 2-norm and squaring on both side, we get

$$\begin{aligned}\left\|\delta^t\right\|^2 &= \left\|\beta\delta^{t-1} + (1-\beta)\left(\overline{\tilde{g}}_{\mathcal{H}}^t - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right) + \beta\left(\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right)\right\|^2\\ &= \beta^2\left\|\delta^{t-1}\right\|^2 + (1-\beta)^2\left\|\overline{\tilde{g}}_{\mathcal{H}}^t - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2 + \beta^2\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\\ &\quad + 2\beta(1-\beta)\left\langle\delta^{t-1},\, \overline{\tilde{g}}_{\mathcal{H}}^t - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle + 2\beta^2\left\langle\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right),\, \delta^{t-1}\right\rangle\\ &\quad + 2\beta(1-\beta)\left\langle\overline{\tilde{g}}_{\mathcal{H}}^t - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right),\, \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle\end{aligned}$$

Taking conditional expectation over the randomness due to sparsification $\mathbb{E}_k\left[\cdot\right]$ on both sides:

$$\begin{aligned}\mathbb{E}_k\left[\left\|\delta^t\right\|^2\right] &= \beta^2\left\|\delta^{t-1}\right\|^2 + (1-\beta)^2\mathbb{E}_k\left[\left\|\overline{\tilde{g}}_{\mathcal{H}}^t - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right] + \beta^2\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\\ &\quad + 2\beta(1-\beta)\left\langle\delta^{t-1},\, \mathbb{E}_k\left[\overline{\tilde{g}}_{\mathcal{H}}^t\right] - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle + 2\beta^2\left\langle\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right),\, \delta^{t-1}\right\rangle\\ &\quad + 2\beta(1-\beta)\left\langle\mathbb{E}_k\left[\overline{\tilde{g}}_{\mathcal{H}}^t\right] - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right),\, \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle\end{aligned}$$

Note that $\mathbb{E}_k\left[\bar{\tilde{g}}_{\mathcal{H}}^t\right] = \mathbb{E}_k\left[\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\tilde{g}_i^t\right] = \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}_k\left[\tilde{g}_i^t\right] = \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}g_i^t = \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\nabla\mathcal{L}_i\left(\theta^{t-1}\right) = \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)$. Substituting above, we get

$$\begin{aligned}
\mathbb{E}_k\left[\left\|\delta^t\right\|^2\right] &\leq \beta^2\left\|\delta^{t-1}\right\|^2 + (1-\beta)^2\mathbb{E}_k\left[\left\|\bar{\tilde{g}}_{\mathcal{H}}^t - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right] + \beta^2\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2 \\
&\quad + 2\beta(1-\beta)\left\langle\delta^{t-1},\, \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle + 2\beta^2\left\langle\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right),\, \delta^{t-1}\right\rangle \\
&\quad + 2\beta(1-\beta)\left\langle\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right),\, \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle \\
&\leq \beta^2\left\|\delta^{t-1}\right\|^2 + (1-\beta)^2\mathbb{E}_k\left[\left\|\bar{\tilde{g}}_{\mathcal{H}}^t - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right] + \beta^2\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2 \\
&\quad + 2\beta^2\left\langle\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right),\, \delta^{t-1}\right\rangle
\end{aligned}$$

Under global sparsification, using Lemma B.1 above we obtain that

$$\mathbb{E}_k\left[\left\|\delta^t\right\|^2\right] \leq (1+\gamma L)\beta^2\left\|\delta^{t-1}\right\|^2 + (1-\beta)^2\left(\tfrac{d}{k}-1\right)\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2 + \gamma L(1+\gamma L)\beta^2\left\|R^{t-1}\right\|^2.$$

Adding and subtracting $\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right)$ from $\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)$ term above, we get

$$\begin{aligned}
\mathbb{E}_k\left[\left\|\delta^t\right\|^2\right] &\leq (1+\gamma L)\beta^2\left\|\delta^{t-1}\right\|^2 + (1-\beta)^2\left(\tfrac{d}{k}-1\right)\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right) + \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right)\right\|^2 \\
&\quad + \gamma L(1+\gamma L)\beta^2\left\|R^{t-1}\right\|^2.
\end{aligned}$$

Using $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$,

$$\begin{aligned}
\mathbb{E}_k\left[\left\|\delta^t\right\|^2\right] &\leq (1+\gamma L)\beta^2\left\|\delta^{t-1}\right\|^2 + (1-\beta)^2\left(\tfrac{d}{k}-1\right)\left(2\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right)\right\|^2 + 2\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right)\right\|^2\right) \\
&\quad + \gamma L(1+\gamma L)\beta^2\left\|R^{t-1}\right\|^2.
\end{aligned}$$

Next, using the L-smoothness property (from Assumption 1), we have $\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right) - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\| \leq L\left\|\theta_{t-2} - \theta_{t-1}\right\|$. Substituting above, we get

$$\begin{aligned}
\mathbb{E}_k\left[\left\|\delta^t\right\|^2\right] &\leq (1+\gamma L)\beta^2\left\|\delta^{t-1}\right\|^2 + (1-\beta)^2\left(\tfrac{d}{k}-1\right)\left(2L^2\left\|\theta_{t-2} - \theta_{t-1}\right\|^2 + 2\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right)\right\|^2\right) \\
&\quad + \gamma L(1+\gamma L)\beta^2\left\|R^{t-1}\right\|^2.
\end{aligned}$$

Recall from Step 4 of our algorithm that $\theta_{t-2} - \theta_{t-1} = \gamma R^{t-1}$. Using this and substituting $\beta_k^2 \coloneqq 2(1-\beta)^2\left(\frac{d}{k}-1\right)$ above, we obtain that

$$\begin{aligned}
\mathbb{E}_k\left[\left\|\delta^t\right\|^2\right] &\leq (1+\gamma L)\beta^2\left\|\delta^{t-1}\right\|^2 + \beta_k^2\gamma^2L^2\left\|R^{t-1}\right\|^2 + \beta_k^2\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right)\right\|^2 + \gamma L(1+\gamma L)\beta^2\left\|R^{t-1}\right\|^2 \\
&\leq (1+\gamma L)\beta^2\left\|\delta^{t-1}\right\|^2 + \beta_k^2\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right)\right\|^2 + \gamma L\left(\gamma L\beta_k^2 + (1+\gamma L)\beta^2\right)\left\|R^{t-1}\right\|^2.
\end{aligned}$$

Recall from Lemma B.3 that we denote $\xi^t \coloneqq R^t - m_{\mathcal{H}}^t$ implying $R^t \coloneqq \xi^t + m_{\mathcal{H}}^t$, and $\delta^t \coloneqq \overline{m}_{\mathcal{H}}^t - \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)$ implying $\overline{m}_{\mathcal{H}}^t \coloneqq \delta^t + \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)$. Thus, using Cauchy-Schwartz inequality, we have $\left\|R^{t-1}\right\|^2 \leq 2\left\|\xi^{t-1}\right\|^2 + 2\left\|\overline{m}_{\mathcal{H}}^{t-1}\right\|^2 = 2\left\|\xi^{t-1}\right\|^2 + 4\left\|\delta^{t-1}\right\|^2 + 4\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_{t-2})\right\|^2$. Using this above, we obtain that

$$\begin{aligned}
\mathbb{E}_k\left[\left\|\delta^t\right\|^2\right] &\leq (1+\gamma L)\beta^2\left\|\delta^{t-1}\right\|^2 + \beta_k^2\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right)\right\|^2 \\
&\quad + \gamma L\left(\gamma L\beta_k^2 + (1+\gamma L)\beta^2\right)\left(2\left\|\xi^{t-1}\right\|^2 + 4\left\|\delta^{t-1}\right\|^2 + 4\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_{t-2})\right\|^2\right) \\
&= \left(4\gamma^2L^2\beta_k^2 + (1+4\gamma L)(1+\gamma L)\beta^2\right)\left\|\delta^{t-1}\right\|^2 + 2\gamma L\left(\gamma L\beta_k^2 + (1+\gamma L)\beta^2\right)\left\|\xi^{t-1}\right\|^2 \\
&\quad + \left((4\gamma^2L^2+1)\beta_k^2 + 4\gamma L(1+\gamma L)\beta^2\right)\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right)\right\|^2.
\end{aligned}$$

Recall that $\beta_k^2 \coloneqq 2(1-\beta)^2\left(\frac{d}{k}-1\right)$, then substituting $\zeta = \left(4\gamma^2L^2\beta_k^2 + (1+4\gamma L)(1+\gamma L)\beta^2\right) = \left(8\gamma^2L^2(1-\beta)^2\left(\frac{d}{k}-1\right) + (1+4\gamma L)(1+\gamma L)\beta^2\right)$ and β_k above, we obtain that

$$\mathbb{E}_k\left[\left\|\delta^t\right\|^2\right] \leq \zeta\left\|\delta^{t-1}\right\|^2 + 2\gamma L\left(2\gamma L(1-\beta)^2\left(\tfrac{d}{k}-1\right) + (1+\gamma L)\beta^2\right)\left\|\xi^{t-1}\right\|^2$$

$$+ \left(2(4\gamma^2L^2+1)(1-\beta)^2\left(\tfrac{d}{k}-1\right)+4\gamma L(1+\gamma L)\beta^2\right)\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-2}\right)\right\|^2.$$

Since t above is an arbitrary value in $[T]$ greater than 1, upon taking expectation over all time steps, on both the sides proves the lemma. ■

We use Lemma 9 from [Allouah et al., 2023a] in our analysis for proving Theorem 1. We restate it below for completeness.

Lemma B.5. *Suppose Assumption 1 holds true. For all $t \in [T]$, we obtain that:*

$$\mathbb{E}\left[2\mathcal{L}_{\mathcal{H}}(\theta_t)-2\mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right] \leq -\gamma(1-4\gamma L)\,\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right]+2\gamma(1+2\gamma L)\,\mathbb{E}\left[\left\|\delta^t\right\|^2\right] + 2\gamma(1+\gamma L)\,\mathbb{E}\left[\left\|\xi^t\right\|^2\right].$$

Proof. Consider an arbitrary step t. Recall from Assumption 1 that $\mathcal{L}_{\mathcal{H}}$ is an L-Lipshitz smooth function, hence we have:

$$\mathcal{L}_{\mathcal{H}}(\theta_t)-\mathcal{L}_{\mathcal{H}}(\theta_{t-1}) \leq \left\langle \theta_t-\theta_{t-1},\, \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle + \tfrac{L}{2}\left\|\theta_t-\theta_{t-1}\right\|^2.$$

Next, recall from our algorithm 1 that $\theta_t=\theta_{t-1}-\gamma R^t$, and recall from Lemma B.3 that we denote $\xi^t=R^t-m^t_{\mathcal{H}}$ implying $R^t=\xi^t+m^t_{\mathcal{H}}$. Thus, we have $\theta_t=\theta_{t-1}-\gamma(\xi^t+m^t_{\mathcal{H}})$. On substituting above, we get:

$$\begin{aligned}\mathcal{L}_{\mathcal{H}}(\theta_t)-\mathcal{L}_{\mathcal{H}}(\theta_{t-1}) &\leq \left\langle -\gamma(\xi^t+m^t_{\mathcal{H}}),\, \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle + \tfrac{L}{2}\left\|\gamma(\xi^t+m^t_{\mathcal{H}})\right\|^2\\ &= -\gamma\left\langle \xi^t,\, \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle - \gamma\left\langle m^t_{\mathcal{H}},\, \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle + \gamma^2\tfrac{L}{2}\left\|\xi^t+m^t_{\mathcal{H}}\right\|^2.\end{aligned}$$

Adding and subtracting $\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)$ from $m^t_{\mathcal{H}}$ in the second term above yields

$$\begin{aligned}\mathcal{L}_{\mathcal{H}}(\theta_t)-\mathcal{L}_{\mathcal{H}}(\theta_{t-1}) \leq -\gamma\left\langle \xi^t,\, \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle &- \gamma\left\langle m^t_{\mathcal{H}}-\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)+\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right),\, \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle\\ &+ \gamma^2\tfrac{L}{2}\left\|\xi^t+m^t_{\mathcal{H}}\right\|^2.\end{aligned}$$

Recall from Lemma B.4 that $\delta^t=m^t_{\mathcal{H}}-\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)$, hence on substituting above we obtain

$$\begin{aligned}\mathcal{L}_{\mathcal{H}}(\theta_t)-\mathcal{L}_{\mathcal{H}}(\theta_{t-1}) &\leq -\gamma\left\langle \xi^t,\, \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle - \gamma\left\langle \delta^t+\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right),\, \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle + \gamma^2\tfrac{L}{2}\left\|\xi^t+m^t_{\mathcal{H}}\right\|^2\\ &= -\gamma\left\langle \xi^t,\, \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle - \gamma\left\langle \delta^t,\, \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle - \gamma\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2 + \gamma^2\tfrac{L}{2}\left\|\xi^t+m^t_{\mathcal{H}}\right\|^2.\end{aligned}$$

Scaling the above equation by 2 yields:

$$2\mathcal{L}_{\mathcal{H}}(\theta_t)-2\mathcal{L}_{\mathcal{H}}(\theta_{t-1}) \leq -2\gamma\left\langle \xi^t,\, \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle - 2\gamma\left\langle \delta^t,\, \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle - 2\gamma\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2 + 2\gamma^2\tfrac{L}{2}\left\|\xi^t+m^t_{\mathcal{H}}\right\|^2. \tag{7}$$

Using Cauchy-Schwartz inequality, and the fact that $2ab \leq \frac{1}{c}a^2+cb^2$ for any $c>0$, we obtain the following inequalities:

$$2\left|\left\langle \delta^t,\, \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle\right| \leq 2\left\|\delta^t\right\|\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\| \leq \tfrac{2}{1}\left\|\delta^t\right\|^2+\tfrac{1}{2}\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2. \tag{8}$$

Similarly,

$$2\left|\left\langle \xi^t,\, \nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\rangle\right| \leq 2\left\|\xi^t\right\|\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\| \leq \tfrac{2}{1}\left\|\xi^t\right\|^2+\tfrac{1}{2}\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2. \tag{9}$$

Finally, using the triangle inequality, we obtain

$$\begin{aligned}\left\|m^t_{\mathcal{H}}+\xi^t\right\| \leq 2\left\|m^t_{\mathcal{H}}\right\|^2+2\left\|\xi^t\right\|^2 &= 2\left\|m^t_{\mathcal{H}}-\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)+\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2+2\left\|\xi^t\right\|^2\\ &\leq 4\left\|m^t_{\mathcal{H}}-\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2+4\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2+2\left\|\xi^t\right\|^2.\end{aligned}$$

Substituting $\delta^t=m^t_{\mathcal{H}}-\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)$ above, we get

$$\left\|m^t_{\mathcal{H}}+\xi^t\right\| \leq 4\left\|\delta^t\right\|^2+4\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2+2\left\|\xi^t\right\|^2. \tag{10}$$

Substituting (8), (9) and (10) in (7), we get

$$2\mathcal{L}_{\mathcal{H}}(\theta_t)-2\mathcal{L}_{\mathcal{H}}(\theta_{t-1}) \leq \gamma\left(2\left\|\xi^t\right\|^2+\tfrac{1}{2}\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right)+\gamma\left(2\left\|\delta^t\right\|^2+\tfrac{1}{2}\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right)$$

$$- 2\gamma \left\|\nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2 + 2\gamma^2 \tfrac{L}{2}\left(4\left\|\delta^t\right\|^2 + 4\left\|\nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2 + 2\left\|\xi^t\right\|^2\right).$$

Rearranging the R.H.S. in the above yields:

$$2\mathcal{L}_{\mathcal{H}}(\theta_t) - 2\mathcal{L}_{\mathcal{H}}(\theta_{t-1}) \leq -\gamma(1-4\gamma L)\left\|\nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2 + 2\gamma(1+2\gamma L)\left\|\delta^t\right\|^2 + 2\gamma(1+\gamma L)\left\|\xi^t\right\|^2.$$

As t is arbitrarily chosen from $[T]$, taking expectation on both sides above proves the lemma. ■

We are now ready to prove Theorem 1, recalled below.

Theorem 1. *Under Assumptions 1 and 2, Algorithm 1 with an (f,κ)-robust aggregation rule F such that $\kappa B^2 \leq \frac{1}{7}$, a learning rate $\gamma \leq \frac{k}{dcL}$ (with $c = 23200$), and a momentum coefficient $\beta = \sqrt{1-24\gamma L}$, satisfies the following guarantee*

$$\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\hat{\theta})\right\|^2\right] \leq \frac{45\left(\mathcal{L}_{\mathcal{H}}(\theta^0) - \mathcal{L}^*_{\mathcal{H}}\right)}{\gamma T(1-\kappa B^2)} + \frac{216\kappa\, G^2}{1-\kappa B^2},$$

where $\mathcal{L}^*_{\mathcal{H}} = \min_{\theta\in\mathbb{R}^d} \mathcal{L}_{\mathcal{H}}(\theta)$.

Proof. Fix an arbitrary $t \in [T]$. We define the Lyapunov function for convergence analysis to be

$$V^t \coloneqq \mathbb{E}\left[2\mathcal{L}_{\mathcal{H}}(\theta_{t-1}) + z\left\|\delta^t\right\|^2 + z'\Upsilon^{t-1}_{\mathcal{H}}\right], \tag{11}$$

where $z = \frac{1}{8L}$ and $z' = \frac{\kappa}{4L}$.

Using the upper bound on $\mathbb{E}\left[\left\|\delta^{t+1}\right\|^2\right]$ from Lemma B.4, we obtain that

$$\begin{aligned}
\mathbb{E}[z\left\|\delta^{t+1}\right\|^2 - z\left\|\delta^t\right\|^2] &\leq z\zeta\,\mathbb{E}\left[\left\|\delta^t\right\|^2\right] + 2z\gamma L\left(2\gamma L(1-\beta)^2\left(\tfrac{d}{k}-1\right) + (1+\gamma L)\beta^2\right)\mathbb{E}\left[\left\|\xi^t\right\|^2\right] \\
&\quad + z\left(2(4\gamma^2L^2+1)(1-\beta)^2\left(\tfrac{d}{k}-1\right) + 4\gamma L(1+\gamma L)\beta^2\right)\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right] - z\,\mathbb{E}\left[\left\|\delta^t\right\|^2\right] \\
&\leq z(\zeta-1)\,\mathbb{E}\left[\left\|\delta^t\right\|^2\right] + 2z\gamma L\left(2\gamma L(1-\beta)^2\left(\tfrac{d}{k}-1\right) + (1+\gamma L)\beta^2\right)\mathbb{E}\left[\left\|\xi^t\right\|^2\right] \\
&\quad + z\left(2(4\gamma^2L^2+1)(1-\beta)^2\left(\tfrac{d}{k}-1\right) + 4\gamma L(1+\gamma L)\beta^2\right)\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right].
\end{aligned} \tag{12}$$

Similarly, using Lemma B.2 we get

$$\begin{aligned}
z'\,\mathbb{E}\left[\Upsilon^t_{\mathcal{H}}\right] - z'\,\mathbb{E}\left[\Upsilon^{t-1}_{\mathcal{H}}\right] &\leq z'\beta\,\mathbb{E}\left[\Upsilon^{t-1}_{\mathcal{H}}\right] + z'\left((1-\beta)^2\tfrac{d}{k} + \beta(1-\beta)\right)\left(G^2 + B^2\left\|\nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|\right) - z'\,\mathbb{E}\left[\Upsilon^{t-1}_{\mathcal{H}}\right] \\
&\leq z'(\beta-1)\,\mathbb{E}\left[\Upsilon^{t-1}_{\mathcal{H}}\right] + z'\left((1-\beta)^2\tfrac{d}{k} + \beta(1-\beta)\right)\left(G^2 + B^2\left\|\nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|\right).
\end{aligned} \tag{13}$$

Similarly, using Lemma B.5, we get

$$\mathbb{E}\left[2\mathcal{L}_{\mathcal{H}}(\theta_t) - 2\mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right] \leq -\gamma(1-4\gamma L)\,\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right] + 2\gamma(1+2\gamma L)\,\mathbb{E}\left[\left\|\delta^t\right\|^2\right] + 2\gamma(1+\gamma L)\,\mathbb{E}\left[\left\|\xi^t\right\|^2\right] \tag{14}$$

Using (11), (12), (13) and (14) below:

$$\begin{aligned}
V^{t+1} - V^t &= \mathbb{E}\left[2\mathcal{L}_{\mathcal{H}}(\theta_t) - 2\mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right] + \mathbb{E}\left[z\left\|\delta^{t+1}\right\|^2 - z\left\|\delta^t\right\|^2\right] + z'\,\mathbb{E}\left[\Upsilon^t_{\mathcal{H}}\right] - z'\,\mathbb{E}\left[\Upsilon^{t-1}_{\mathcal{H}}\right] \\
&\leq -\gamma(1-4\gamma L)\,\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right] + 2\gamma(1+2\gamma L)\,\mathbb{E}\left[\left\|\delta^t\right\|^2\right] + 2\gamma(1+\gamma L)\,\mathbb{E}\left[\left\|\xi^t\right\|^2\right] \\
&\quad + z(\zeta-1)\,\mathbb{E}\left[\left\|\delta^t\right\|^2\right] + 2z\gamma L\left(2\gamma L(1-\beta)^2\left(\tfrac{d}{k}-1\right) + (1+\gamma L)\beta^2\right)\mathbb{E}\left[\left\|\xi^t\right\|^2\right] \\
&\quad + z\left(2(4\gamma^2L^2+1)(1-\beta)^2\left(\tfrac{d}{k}-1\right) + 4\gamma L(1+\gamma L)\beta^2\right)\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right] \\
&\quad + z'(\beta-1)\,\mathbb{E}\left[\Upsilon^{t-1}_{\mathcal{H}}\right] + z'\left((1-\beta)^2\tfrac{d}{k} + \beta(1-\beta)\right)\left(G^2 + B^2\left\|\nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|\right)
\end{aligned}$$

$$
\begin{aligned}
&= -\gamma \Big(1 - 4\gamma L - \tfrac{z}{\gamma}\left(2(4\gamma^2L^2+1)(1-\beta)^2\left(\tfrac{d}{k}-1\right) + 4\gamma L(1+\gamma L)\beta^2\right) \\
&\quad - \tfrac{z'}{\gamma}\left((1-\beta)^2\tfrac{d}{k} + \beta(1-\beta)\right)B^2\Big)\,\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^t\right)\right\|^2\right] + \left(2\gamma(1+2\gamma L) + z(\zeta-1)\right)\mathbb{E}\left[\left\|\delta^t\right\|^2\right] \\
&\quad + 2\gamma\left(1+\gamma L + zL\left(2\gamma L(1-\beta)^2\left(\tfrac{d}{k}-1\right) + (1+\gamma L)\beta^2\right)\right)\mathbb{E}\left[\left\|\xi^t\right\|^2\right] + z'(\beta-1)\,\mathbb{E}\left[\Upsilon_{\mathcal{H}}^{t-1}\right] \\
&\quad + z'\left((1-\beta)^2\tfrac{d}{k} + \beta(1-\beta)\right)G^2
\end{aligned} \tag{15}
$$

Denote

$$
\begin{aligned}
&A_1 = 1 - 4\gamma L - \tfrac{z}{\gamma}\left(2(4\gamma^2L^2+1)(1-\beta)^2\left(\tfrac{d}{k}-1\right) + 4\gamma L(1+\gamma L)\beta^2\right) - \tfrac{z'}{\gamma}\left((1-\beta)^2\tfrac{d}{k} + \beta(1-\beta)\right)B^2, \\
&A_2 = 2\gamma(1+2\gamma L) + z(\zeta-1), \quad \text{and} \\
&A_3 = 1 + \gamma L + zL\left(2\gamma L(1-\beta)^2\left(\tfrac{d}{k}-1\right) + (1+\gamma L)\beta^2\right).
\end{aligned} \tag{16}
$$

Then, we can rewrite (15) as:

$$
\begin{aligned}
V^{t+1} - V^t \le &-\gamma A_1\,\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right] + A_2\,\mathbb{E}\left[\left\|\delta^t\right\|^2\right] + 2\gamma A_3\,\mathbb{E}\left[\left\|\xi^t\right\|^2\right] + z'(\beta-1)\,\mathbb{E}\left[\Upsilon_{\mathcal{H}}^{t-1}\right] \\
&+ z'\left((1-\beta)^2\tfrac{d}{k} + \beta(1-\beta)\right)G^2.
\end{aligned} \tag{17}
$$

Consider term A_1. Since $\beta < 1$, we obtain that

$$
A_1 \ge 1 - 4\gamma L - \tfrac{z}{\gamma}\left(2(4\gamma^2L^2+1)(1-\beta^2)^2\left(\tfrac{d}{k}-1\right) + 4\gamma L(1+\gamma L)\beta^2\right) - \tfrac{z'}{\gamma}\left(\left(1-\beta^2\right)^2\tfrac{d}{k} + \tfrac{1}{2}(1-\beta^2)\right)B^2.
$$

Substituting $1-\beta^2 = 24\gamma L$ above, we obtain that

$$
\begin{aligned}
A_1 &\ge 1 - 4\gamma L - \tfrac{z}{\gamma}\left(2(4\gamma^2L^2+1)(24\gamma L)^2\left(\tfrac{d}{k}-1\right) + 4\gamma L(1+\gamma L)(1-24\gamma L)\right) - \tfrac{z'}{\gamma}(24\gamma L)^2\tfrac{d}{k}B^2 - \tfrac{z'}{\gamma}12\gamma L)B^2 \\
&= 1 - 4\gamma L - 2\gamma z(4\gamma^2L^2+1)(24L)^2\left(\tfrac{d}{k}-1\right) - 4zL(1+\gamma L)(1-24\gamma L) - 576z'\gamma L^2\tfrac{d}{k}B^2 - 12z'LB^2.
\end{aligned}
$$

Substituting $z = \frac{1}{8L}$ and $z' = \frac{\kappa}{4L}$ above yields,

$$
\begin{aligned}
A_1 &\ge 1 - 4\gamma L - \tfrac{2\gamma}{8L}(4\gamma^2L^2+1)(24L)^2\left(\tfrac{d}{k}-1\right) - \tfrac{4L}{8L}(1+\gamma L) + 12\gamma L + 12(\gamma L)^2 - \left(\tfrac{144\kappa}{L}\right)\gamma L^2\tfrac{d}{k}B^2 - \left(\tfrac{3\kappa}{L}\right)LB^2 \\
&= \tfrac{1}{2} + \tfrac{15}{2}\gamma L + 12(\gamma L)^2 - 144\gamma L(4\gamma^2L^2+1)\left(\tfrac{d}{k}-1\right) - 144\kappa\gamma L\tfrac{d}{k}B^2 - 3\kappa B^2 \\
&\ge \tfrac{1}{2} - 144\gamma L(4\gamma^2L^2+1)\left(\tfrac{d}{k}-1\right) - 144\kappa\gamma L\tfrac{d}{k}B^2 - 3\kappa B^2.
\end{aligned}
$$

Recall that $\gamma \le \frac{k/d}{cL}$ where $c = 23200$. Moreover, we assume that $\kappa \le \frac{1}{7B^2}$. Using this above yields,

$$
\begin{aligned}
A_1 &\ge \tfrac{1}{2} - 144L\left(\tfrac{d}{k}-1\right)\times\tfrac{k/d}{23200L}\left(1 + \tfrac{(k/d)^2}{23200^2}\right) - 144L\tfrac{d}{k}B^2\times\tfrac{k/d}{23200L}\times\tfrac{1}{7B^2} - \tfrac{3B^2}{7B^2} \\
&= \tfrac{1}{2} - \tfrac{144}{23200}\left(1-\tfrac{k}{d}\right)\times\left(1 + \tfrac{(k/d)^2}{23200^2}\right) - \tfrac{144}{23200}\times\tfrac{1}{7} - \tfrac{3}{7} \\
&\ge \tfrac{1}{2} - \tfrac{1}{160}\left(1 + \tfrac{1}{23200^2}\right) - \tfrac{1}{160\times7} - \tfrac{3}{7} > \tfrac{1}{18}.
\end{aligned} \tag{18}
$$

Consider term A_2. Substituting $\zeta = 8\gamma^2L^2(1-\beta)^2\left(\frac{d}{k}-1\right) + (1+4\gamma L)(1+\gamma L)\beta^2$, we obtain that

$$
\begin{aligned}
A_2 &= 2\gamma(1+2\gamma L) + z(\zeta-1) \\
&= 2\gamma(1+2\gamma L) + z\left(8\gamma^2L^2(1-\beta)^2\left(\tfrac{d}{k}-1\right) + (1+4\gamma L)(1+\gamma L)\beta^2 - 1\right) \\
&= 2\gamma(1+2\gamma L) - z(1-\beta^2) + 8z\gamma^2L^2(1-\beta^2)^2\left(\tfrac{d}{k}-1\right) + z\gamma L(5+4\gamma L)\beta^2.
\end{aligned}
$$

Since $\beta < 1$, from above we obtain that

$$
A_2 \le 2\gamma(1+2\gamma L) - z(1-\beta^2) + 8z\gamma^2L^2(1-\beta^2)^2\left(\tfrac{d}{k}-1\right) + z\gamma L(5+4\gamma L).
$$

Substituting $1-\beta^2 = 24\gamma L$ above, we obtain that

$$A_2 \leq -z(24\gamma L) + 8z\gamma^2 L^2 (24\gamma L)^2 \left(\tfrac{d}{k} - 1\right) + z\gamma L(5 + 4\gamma L) + 2\gamma(1 + 2\gamma L).$$

Substituting $z = \frac{1}{8L}$ above yields,

$$\begin{aligned} A_2 &\leq -\tfrac{24\gamma L}{8L} + \gamma \left(\tfrac{8\gamma L^2}{8L} (24\gamma L)^2 \left(\tfrac{d}{k} - 1\right) + \tfrac{L}{8L}(5 + 4\gamma L) + 2(1 + 2\gamma L) \right) \\ &\leq -3\gamma + \gamma \left(\gamma L (24\gamma L)^2 \left(\tfrac{d}{k} - 1\right) + \tfrac{1}{8}(5 + 4\gamma L) + 2(1 + 2\gamma L) \right). \end{aligned}$$

Since $\gamma \leq \frac{1}{24L}$, from above we obtain that

$$\begin{aligned} A_2 &\leq -3\gamma + \gamma \left(\gamma L (\tfrac{24L}{24L})^2 \left(\tfrac{d}{k} - 1\right) + \tfrac{1}{8}(5 + \tfrac{4L}{24L}) + 2(1 + \tfrac{2L}{24L}) \right) \\ &\leq -3\gamma + \gamma \left(\gamma L \left(\tfrac{d}{k} - 1\right) + \tfrac{1}{8}(\tfrac{31}{6}) + \tfrac{13}{6} \right) \\ &= -3\gamma + \gamma \left(\left(\tfrac{d}{k} - 1\right) \gamma L + \tfrac{45}{16} \right). \end{aligned}$$

Since $\gamma \leq \frac{k/d}{23200L}$, the above yields,

$$A_2 \leq -3\gamma + \gamma \left(\left(\tfrac{d}{k} - 1\right) L \times \tfrac{k/d}{23200L} + \tfrac{45}{16} \right) < -3\gamma + \tfrac{46}{16}\gamma < 0. \tag{19}$$

Consider term A_3. Since $\beta < 1$, we obtain that

$$\begin{aligned} A_3 &= 1 + \gamma L + zL \left(2\gamma L (1-\beta^2)^2 \tfrac{d}{k} + (1 + \gamma L)\beta^2 \right) \\ &\leq 1 + \gamma L + zL \left(2\gamma L (1-\beta^2)^2 \tfrac{d}{k} + (1 + \gamma L) \right). \end{aligned}$$

Substituting $1 - \beta^2 = 24\gamma L$, we obtain that

$$A_3 \leq 1 + \gamma L + zL \left(2\gamma L (24\gamma L)^2 \left(\tfrac{d}{k} - 1\right) + (1 + \gamma L) \right).$$

Substituting $z = \frac{1}{8L}$ above, we get

$$A_3 \leq 1 + \gamma L + \tfrac{L}{8L} \left(2\gamma L (24\gamma L)^2 \left(\tfrac{d}{k} - 1\right) + (1 + \gamma L) \right).$$

Since $\gamma \leq \frac{1}{24L}$, from above we obtain that

$$\begin{aligned} A_3 &\leq 1 + \tfrac{L}{24L} + \tfrac{\gamma L}{4} (\tfrac{24L}{24L})^2 \left(\tfrac{d}{k} - 1\right) + \tfrac{1}{8}(1 + \tfrac{L}{24L}) = \tfrac{25}{24} + \tfrac{1}{4}\gamma L \left(\tfrac{d}{k} - 1\right) + \tfrac{1}{8}(\tfrac{25}{24}) \\ &\leq \tfrac{75}{64} + \tfrac{1}{4} \left(\tfrac{d}{k} - 1\right) \gamma L. \end{aligned}$$

Recalling that $\gamma \leq \frac{k/d}{23200L}$, we obtain,

$$A_3 \leq \tfrac{75}{64} + \tfrac{1}{4} \left(\tfrac{d}{k} - 1\right) L \times \tfrac{k/d}{23200L} = \tfrac{75}{64} + \tfrac{1}{92800} \left(1 - \tfrac{k}{d}\right) < \tfrac{76}{64} < \tfrac{3}{2}. \tag{20}$$

Substituting values from (18), (19) and (20) in (17):

$$\begin{aligned} V^{t+1} - V^t \leq & -\tfrac{\gamma}{18} \mathbb{E} \left[\left\| \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t-1}) \right\|^2 \right] + 3\gamma \mathbb{E} \left[\left\| \xi^t \right\|^2 \right] + z'(\beta - 1) \mathbb{E} \left[\Upsilon_{\mathcal{H}}^{t-1} \right] \\ & + z' \left((1-\beta)^2 \tfrac{d}{k} + \beta(1-\beta) \right) G^2. \end{aligned}$$

Substituting value of $\mathbb{E}\left[\left\|\xi^t\right\|^2\right]$ from Lemma B.3 above, we get

$$\begin{aligned} V^{t+1} - V^t \leq & -\tfrac{\gamma}{18} \mathbb{E} \left[\left\| \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t-1}) \right\|^2 \right] + 3\gamma\kappa \Big(\beta \mathbb{E} \left[\Upsilon_{\mathcal{H}}^{t-1} \right] + \left((1-\beta)^2 \tfrac{d}{k} + \beta(1-\beta) \right) \Big(G^2 \\ & + B^2 \mathbb{E} \left[\left\| \nabla \mathcal{L}_{\mathcal{H}} \left(\theta^{t-1} \right) \right\|^2 \right] \Big) \Big) + z'(\beta - 1) \mathbb{E} \left[\Upsilon_{\mathcal{H}}^{t-1} \right] + z' \left((1-\beta)^2 \tfrac{d}{k} + \beta(1-\beta) \right) G^2 \\ \leq & -\left(\tfrac{\gamma}{18} - 3\gamma\kappa \left((1-\beta)^2 \tfrac{d}{k} + \beta(1-\beta) \right) B^2 \right) \mathbb{E} \left[\left\| \nabla \mathcal{L}_{\mathcal{H}}(\theta_{t-1}) \right\|^2 \right] + (3\gamma\kappa + z') \end{aligned}$$

$$\times \left(\left(1-\beta\right)^2 \tfrac{d}{k} + \beta(1-\beta)\right) G^2 + (3\gamma\kappa\beta - (1-\beta)z')\,\mathbb{E}\left[\Upsilon_{\mathcal{H}}^{t-1}\right].$$

Since $1-\beta = \frac{1-\beta^2}{1+\beta}$, we have $\frac{1-\beta^2}{2} \leq 1-\beta \leq 1-\beta^2$, resulting in

$$\begin{aligned} V^{t+1} - V^t \leq & -\tfrac{\gamma}{18}\left(1 - 54\kappa\left(\left(1-\beta^2\right)^2 \tfrac{d}{k} + \beta(1-\beta^2)\right) B^2\right) \mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right] \\ & + \left(3\gamma\kappa\beta - \tfrac{(1-\beta^2)}{2} z'\right) \mathbb{E}\left[\Upsilon_{\mathcal{H}}^{t-1}\right] + (3\gamma\kappa + z')\left(\left(1-\beta^2\right)^2 \tfrac{d}{k} + \beta(1-\beta^2)\right) G^2. \end{aligned}$$

Substituting $1-\beta^2 = 24\gamma L$ above, we obtain

$$\begin{aligned} V^{t+1} - V^t &\leq -\tfrac{\gamma}{18}\left(1 - 54\kappa\left(\left(24\gamma L\right)^2 \tfrac{d}{k} + 24\gamma L\right) B^2\right) \mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right] \\ &\quad + (3\gamma\kappa + z')\left(\left(24\gamma L\right)^2 \tfrac{d}{k} + 24\gamma L\right) G^2 + \left(3\gamma\kappa - \tfrac{24\gamma}{2} z' L\right) \mathbb{E}\left[\Upsilon_{\mathcal{H}}^{t-1}\right] \\ &= -\tfrac{\gamma}{18}\left(1 - 54\kappa\left(\left(24\gamma L\right)^2 \tfrac{d}{k} + 24\gamma L\right) B^2\right) \mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right] \\ &\quad + (3\gamma\kappa + z')\left(576\gamma^2 L^2 \tfrac{d}{k} + 24\gamma L\right) G^2 + (3\gamma\kappa - 12\gamma z' L)\,\mathbb{E}\left[\Upsilon_{\mathcal{H}}^{t-1}\right]. \end{aligned}$$

Since $\gamma \leq \frac{1}{24L}$, from above we obtain that

$$\begin{aligned} V^{t+1} - V^t &\leq -\tfrac{\gamma}{18}\left(1 - 54\kappa\left(\left(\tfrac{24L}{24L} \times 24\gamma L\right) \tfrac{d}{k} + 24\gamma L\right) B^2\right) \mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right] \\ &\quad + (3\gamma\kappa - 12z'\gamma L)\,\mathbb{E}\left[\Upsilon_{\mathcal{H}}^{t-1}\right] + (3\kappa\gamma + z')\left(576\gamma^2 L^2 \tfrac{d}{k} + 24\gamma L\right) G^2 \\ &= -\tfrac{\gamma}{18}\left(1 - 54\kappa\left(24\gamma L \tfrac{d}{k} + 24\gamma L\right) B^2\right) \mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right] \\ &\quad + (3\gamma\kappa - 12z'\gamma L)\,\mathbb{E}\left[\Upsilon_{\mathcal{H}}^{t-1}\right] + (3\kappa\gamma + z')\left(576\gamma^2 L^2 \tfrac{d}{k} + 24\gamma L\right) G^2. \end{aligned}$$

Substituting $z' = \frac{\kappa}{4L}$ above, we obtain that

$$\begin{aligned} V^{t+1} - V^t &\leq -\tfrac{\gamma}{18}\left(1 - 54\kappa\left(24\gamma L \tfrac{d}{k} + 24\gamma L\right) B^2\right) \mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right] \\ &\quad + \left(3\gamma\kappa - \tfrac{12L\kappa}{4L}\gamma\right) \mathbb{E}\left[\Upsilon_{\mathcal{H}}^{t-1}\right] + \left(3\kappa\gamma + \tfrac{\kappa}{4L}\right)\left(576\gamma^2 L^2 \tfrac{d}{k} + 24\gamma L\right) G^2 \\ &= -\tfrac{\gamma}{18}\left(1 - 54\kappa\left(24\gamma L \tfrac{d}{k} + 24\gamma L\right) B^2\right) \mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right] \\ &\quad + (3\gamma\kappa - 3\kappa\gamma)\,\mathbb{E}\left[\Upsilon_{\mathcal{H}}^{t-1}\right] + \tfrac{\kappa}{4L}\left(12\gamma L + 1\right)\left(576\gamma^2 L^2 \tfrac{d}{k} + 24\gamma L\right) G^2 \\ &= -\tfrac{\gamma}{18}\left(1 - 54\kappa\left(24\gamma L \tfrac{d}{k} + 24\gamma L\right) B^2\right) \mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right] + \kappa\left(12\gamma L + 1\right)\left(144\gamma^2 L \tfrac{d}{k} + 6\gamma\right) G^2. \end{aligned}$$

Substituting $\gamma \leq \frac{k/d}{23200L} \leq \frac{1}{23200L}$ yields,

$$\begin{aligned} V^{t+1} - V^t &\leq -\tfrac{\gamma}{18}\left(1 - 54\kappa\left(24L \tfrac{d}{k} \times \tfrac{k/d}{23200L} + \tfrac{24L}{23200L}\right) B^2\right) \mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right] \\ &\quad + \kappa\left(12\gamma L + 1\right)\left(144\gamma^2 L \tfrac{d}{k} + 6\gamma\right) G^2 \\ &< -\tfrac{\gamma}{18}\left(1 - \tfrac{54\times 48}{23200}\kappa B^2\right) \mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right] + 6\kappa\gamma\left(288\gamma^2 L^2 \tfrac{d}{k} + 24\gamma L \tfrac{d}{k} + 12\gamma L + 1\right) G^2. \end{aligned}$$

Note that $\left(1 - \frac{54\times 48}{23200}\kappa B^2\right) \geq 1 - \kappa B^2$, which is positive since $7\kappa B^2 \leq 1$. Using this above yields,

$$V^{t+1} - V^t \leq -\tfrac{\gamma}{18}(1 - \kappa B^2)\,\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right] + 6\kappa\gamma\left(288\gamma^2 L^2 \tfrac{d}{k} + 24\gamma L \tfrac{d}{k} + 12\gamma L + 1\right) G^2.$$

Moving the term $\nabla \mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)$ to the L.H.S, we get

$$\tfrac{\gamma}{18}(1 - \kappa B^2)\,\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right] \leq V^t - V^{t+1} + 6\kappa\gamma\left(288\gamma^2 L^2 \tfrac{d}{k} + 24\gamma L \tfrac{d}{k} + 12\gamma L + 1\right) G^2.$$

Taking summation on both the sides from $t=1$ to T and rearranging, we get

$$\tfrac{\gamma(1-\kappa B^2)}{18}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right] \leq V^1 - V^{T+1} + 6\kappa\gamma\left(288\gamma^2L^2\tfrac{d}{k} + 24\gamma L\tfrac{d}{k} + 12\gamma L + 1\right)G^2T.$$

Multiplying both sides by $\frac{18}{\gamma(1-\kappa B^2)}$,

$$\textstyle\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right] \leq \frac{18\left(V^1-V^{T+1}\right)}{\gamma(1-\kappa B^2)} + 108\kappa\left(288\gamma^2L^2\frac{d}{k} + 24\gamma L\frac{d}{k} + 12\gamma L + 1\right)\frac{G^2T}{1-\kappa B^2}.$$

Dividing both sides by T, we get,

$$\tfrac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right] \leq \frac{18\left(V^1-V^{T+1}\right)}{\gamma T(1-\kappa B^2)} + 108\kappa\left(288\gamma^2L^2\frac{d}{k} + 24\gamma L\frac{d}{k} + 12\gamma L + 1\right)\frac{G^2}{1-\kappa B^2}. \tag{21}$$

Next, we obtain an upper bound on $V^1 - V^{T+1}$. Let $\mathcal{L}^*_{\mathcal{H}} = \inf_{\theta\in\mathbb{R}^d}\mathcal{L}_{\mathcal{H}}(\theta)$. For any $t>0$, using definition of V^t from (11), we have:

$$V^t - 2\mathcal{L}^*_{\mathcal{H}} = 2\,\mathbb{E}\left[\mathcal{L}_{\mathcal{H}}(\theta_{t-1}) - \mathcal{L}^*_{\mathcal{H}}\right] + z\,\mathbb{E}\left[\left\|\delta^t\right\|^2\right] \geq 0 + z\,\mathbb{E}\left[\left\|\delta^t\right\|^2\right] + z'\,\mathbb{E}\left[\Upsilon^{t-1}_{\mathcal{H}}\right] \geq 0.$$

Thus, $V^t \geq 2\mathcal{L}^*_{\mathcal{H}}$, implying:

$$V^1 - V^{T+1} \leq V^1 - 2\mathcal{L}^*_{\mathcal{H}}. \tag{22}$$

Note that for $t=1$, from (11), and the definitions of δ^t and $\Upsilon^0_{\mathcal{H}}$ in Lemmas B.4 and B.2, respectively, we have

$$\begin{aligned}
V^1 &= 2\mathcal{L}_{\mathcal{H}}(\theta_0) + z\,\mathbb{E}\left[\left\|\delta^1\right\|^2\right] + z'\,\mathbb{E}\left[\Upsilon^0_{\mathcal{H}}\right] \\
&\leq 2\mathcal{L}_{\mathcal{H}}(\theta_0) + z\,\mathbb{E}\left[\left\|\overline{m}^1_{\mathcal{H}} - \nabla\mathcal{L}_{\mathcal{H}}(\theta_0)\right\|^2\right] + z'\tfrac{1}{\mathcal{H}}\textstyle\sum_{i\in\mathcal{H}}\mathbb{E}_k\left[\left\|\tilde{m}^0_i\right\|^2\right] \\
&\leq 2\mathcal{L}_{\mathcal{H}}(\theta_0) + z\,\mathbb{E}\left[\left\|\tfrac{1}{\mathcal{H}}\textstyle\sum_{i\in\mathcal{H}}m^1_i - \nabla\mathcal{L}_{\mathcal{H}}(\theta_0)\right\|^2\right] + z'\tfrac{1}{\mathcal{H}}\textstyle\sum_{i\in\mathcal{H}}\mathbb{E}_k\left[\left\|\mathcal{M}^0_i\right\|^2\right].
\end{aligned}$$

Since $m^0_i = 0$, $\mathcal{M}^0_i \coloneqq m^t_i - m^t_{\mathcal{H}} = 0$ for all $i\in\mathcal{H}$. Using this above, we obtain that

$$V^1 \leq 2\mathcal{L}_{\mathcal{H}}(\theta_0) + z\,\mathbb{E}\left[\left\|\overline{m}^1_{\mathcal{H}} - \nabla\mathcal{L}_{\mathcal{H}}(\theta_0)\right\|^2\right] = 2\mathcal{L}_{\mathcal{H}}(\theta_0) + z\,\mathbb{E}\left[\left\|\tfrac{1}{\mathcal{H}}\textstyle\sum_{i\in\mathcal{H}}m^1_i - \nabla\mathcal{L}_{\mathcal{H}}(\theta_0)\right\|^2\right].$$

Recall that $m^t_i = \beta m^{t-1}_i + (1-\beta)\tilde{g}^t_i$ where $m^0_i = 0$. Thus, $m^1_i = (1-\beta)\tilde{g}^1_i$. Using this above, we get

$$\begin{aligned}
V^1 &\leq 2\mathcal{L}_{\mathcal{H}}(\theta_0) + z\,\mathbb{E}\left[\left\|\tfrac{1}{\mathcal{H}}\textstyle\sum_{i\in\mathcal{H}}(1-\beta)\tilde{g}^1_i - \nabla\mathcal{L}_{\mathcal{H}}(\theta_0)\right\|^2\right] \\
&= 2\mathcal{L}_{\mathcal{H}}(\theta_0) + z\,\mathbb{E}\left[\left\|(1-\beta)\overline{\tilde{g}}^1_{\mathcal{H}} - \nabla\mathcal{L}_{\mathcal{H}}(\theta_0)\right\|^2\right] \\
&= 2\mathcal{L}_{\mathcal{H}}(\theta_0) + z\,\mathbb{E}\left[\left\|(1-\beta)\left(\overline{\tilde{g}}^1_{\mathcal{H}} - \nabla\mathcal{L}_{\mathcal{H}}(\theta_0)\right) + \beta\nabla\mathcal{L}_{\mathcal{H}}(\theta_0)\right\|^2\right].
\end{aligned} \tag{23}$$

Note that

$$\mathbb{E}\left[\overline{\tilde{g}}^1_{\mathcal{H}}\right] = \mathbb{E}_t\left[\mathbb{E}_k\left[\overline{\tilde{g}}^1_{\mathcal{H}}\right]\right] = \mathbb{E}_t\left[\overline{g}^1_{\mathcal{H}}\right] = \mathbb{E}_t\left[\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}g^1_i\right] = \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\mathbb{E}_t\left[g^1_i\right] = \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\nabla\mathcal{L}_i\left(\theta^0\right) = \nabla\mathcal{L}_{\mathcal{H}}(\theta_0).$$

Therefore,

$$\mathbb{E}\left[\left\|(1-\beta)\left(\overline{\tilde{g}}^1_{\mathcal{H}} - \nabla\mathcal{L}_{\mathcal{H}}(\theta_0)\right) + \beta\nabla\mathcal{L}_{\mathcal{H}}(\theta_0)\right\|^2\right] = \mathbb{E}\left[\left\|(1-\beta)\left(\overline{\tilde{g}}^1_{\mathcal{H}} - \nabla\mathcal{L}_{\mathcal{H}}(\theta_0)\right)\right\|^2 + \left\|\beta\nabla\mathcal{L}_{\mathcal{H}}(\theta_0)\right\|^2\right].$$

Using above in (23), we get:

$$V^1 \leq 2\mathcal{L}_{\mathcal{H}}(\theta_0) + z\left((1-\beta)^2\,\mathbb{E}\left[\left\|\bar{\tilde{g}}^1_{\mathcal{H}} - \nabla\mathcal{L}_{\mathcal{H}}(\theta_0)\right\|^2\right] + \beta^2\,\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^0\right)\right\|^2\right]\right).$$

Since $\mathbb{E}\left[\left\|\bar{\tilde{g}}^t_{\mathcal{H}} - \nabla\mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right] \leq \left(\frac{d}{k}-1\right)\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^{t-1}\right)\right\|^2\right]$, owing to Lemma B.1, we get:

$$\begin{aligned} V^1 &\leq 2\mathcal{L}_{\mathcal{H}}(\theta_0) + z\left((1-\beta)^2\left(\tfrac{d}{k}-1\right)\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^0\right)\right\|^2\right] + \beta^2\,\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^0\right)\right\|^2\right]\right) \\ &= 2\mathcal{L}_{\mathcal{H}}(\theta_0) + z\left((1-\beta^2)^2\left(\tfrac{d}{k}-1\right) + \beta^2\right)\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^0\right)\right\|^2\right]. \end{aligned}$$

Recall that $\beta^2 < 1$, $1-\beta^2 = 24\gamma L$ and $z = \frac{1}{8L}$. Therefore,

$$\begin{aligned} V^1 &\leq 2\mathcal{L}_{\mathcal{H}}(\theta_0) + \tfrac{1}{8L}\left((24\gamma L)^2\left(\tfrac{d}{k}-1\right)+1\right)\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^0\right)\right\|^2\right] \\ &= 2\mathcal{L}_{\mathcal{H}}(\theta_0) + (72\gamma^2 L\left(\tfrac{d}{k}-1\right) + \tfrac{1}{8L})\,\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta^0\right)\right\|^2\right]. \end{aligned}$$

Due to the L-Lipschitz smoothness (cf. Assumption 1) we have $\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}\left(\theta\right)\right\|^2\right] \leq 2L\left(\mathcal{L}_{\mathcal{H}}(\theta) - \mathcal{L}^*_{\mathcal{H}}\right)$ (See [Nesterov et al., 2018], Theorem 2.1.5). Substituting above, we get

$$\begin{aligned} V^1 &\leq 2\mathcal{L}_{\mathcal{H}}(\theta_0) + (72\gamma^2 L\left(\tfrac{d}{k}-1\right) + \tfrac{1}{8L})2L\left(\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}^*_{\mathcal{H}}\right) \\ &\leq 2\mathcal{L}_{\mathcal{H}}(\theta_0) + (144\gamma^2 L^2\left(\tfrac{d}{k}-1\right) + \tfrac{1}{4})\left(\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}^*_{\mathcal{H}}\right). \end{aligned}$$

Substituting from above in (22), we obtain that

$$\begin{aligned} V^1 - V^{T+1} &\leq 2(\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}^*) + (144\gamma^2 L^2\left(\tfrac{d}{k}-1\right) + \tfrac{1}{4})\left(\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}^*_{\mathcal{H}}\right) \\ &\leq (144\gamma^2 L^2\left(\tfrac{d}{k}-1\right) + \tfrac{9}{4})\left(\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}^*_{\mathcal{H}}\right). \end{aligned}$$

Recall that $\gamma \leq \frac{k/d}{23200L} \leq \frac{1}{23200L}$. Substituting above, we get

$$\begin{aligned} V^1 - V^{T+1} &\leq (\tfrac{144L^2\times k/d}{(23200)^2L^2}\left(\tfrac{d}{k}-1\right) + \tfrac{9}{4})\left(\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}^*_{\mathcal{H}}\right) = (\tfrac{144L^2}{(23200)^2L^2}\left(1-\tfrac{k}{d}\right) + \tfrac{9}{4})\left(\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}^*_{\mathcal{H}}\right) \\ &\leq \tfrac{10}{4}\left(\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}^*_{\mathcal{H}}\right) = \tfrac{5}{2}\left(\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}^*_{\mathcal{H}}\right). \end{aligned}$$

Substituting from above in (21) proves the theorem, i.e., we obtain that

$$\begin{aligned} \tfrac{1}{T}\textstyle\sum_{t=1}^T \mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right] &\leq \tfrac{18\times\frac{5}{2}\left(\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}^*_{\mathcal{H}}\right)}{\gamma T(1-\kappa B^2)} + 108\kappa\left(288\gamma^2L^2\tfrac{d}{k} + 24\gamma L\tfrac{d}{k} + 12\gamma L + 1\right)\tfrac{G^2}{1-\kappa B^2} \\ &= \tfrac{45\left(\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}^*_{\mathcal{H}}\right)}{\gamma T(1-\kappa B^2)} + 108\kappa\left(288\gamma^2L^2\tfrac{d}{k} + 24\gamma L\tfrac{d}{k} + 12\gamma L + 1\right)\tfrac{G^2}{1-\kappa B^2}. \end{aligned}$$

Note that since $\gamma \leq \frac{k/d}{23200L} \leq \frac{1}{23200L}$,

$$288\gamma^2L^2\tfrac{d}{k} + 24\gamma L\tfrac{d}{k} + 12\gamma L \leq \tfrac{288L}{23200}\gamma + \tfrac{24}{23200} + 12\gamma L \leq \tfrac{288}{23200^2} + \tfrac{24}{23200} + \tfrac{12}{23200} < 1$$

Using this above, we obtain that

$$\tfrac{1}{T}\textstyle\sum_{t=1}^T \mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right] \leq \tfrac{45\left(\mathcal{L}_{\mathcal{H}}(\theta_0) - \mathcal{L}^*_{\mathcal{H}}\right)}{\gamma T(1-\kappa B^2)} + \tfrac{216\kappa G^2}{1-\kappa B^2}.$$

Since, $\hat{\theta}$ is chosen uniformly at random from $\{\theta_0, \ldots, \theta_{T-1}\}$, $\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}(\hat{\theta})\right\|^2\right] = \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{H}}(\theta_{t-1})\right\|^2\right]$. Using this above proves the theorem. ■

Corollary 1. *In the same conditions of Theorem 1, let* $\gamma = \frac{1}{\alpha c L}$*, then*

$$\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\hat{\theta})\right\|^2\right] \leq \mathcal{O}\Big(\tfrac{\alpha}{T(1-\kappa B^2)} + \tfrac{\kappa G^2}{1-\kappa B^2}\Big).$$

Proof. Recall from Theorem 1 that

$$\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\hat{\theta})\right\|^2\right] \leq \tfrac{45(\mathcal{L}_{\mathcal{H}}(\theta_0)-\mathcal{L}^*_{\mathcal{H}})}{\gamma T(1-\kappa B^2)} + \tfrac{216\kappa G^2}{1-\kappa B^2}.$$

Also, for $\alpha = \frac{d}{k}$, recall that $\gamma = \frac{1}{c\alpha L}$ implying that $\frac{1}{\gamma} = c\alpha L$, which results in

$$\mathbb{E}\left[\left\|\nabla \mathcal{L}_{\mathcal{H}}(\hat{\theta})\right\|^2\right] \leq \tfrac{45cL\alpha(\mathcal{L}_{\mathcal{H}}(\theta_0)-\mathcal{L}^*_{\mathcal{H}})}{T(1-\kappa B^2)} + \tfrac{216\kappa G^2}{1-\kappa B^2} = \mathcal{O}\left(\tfrac{\alpha}{T(1-\kappa B^2)} + \tfrac{\kappa G^2}{1-\kappa B^2}\right),$$

which concludes the proof. ■

C Description and theoretical guarantees of Byz-DASHA-PAGE

In this section, we discuss the interpretation of the theoretical results of Byz-DASHA-PAGE [Rammal et al., 2024] under the $(f,\kappa)-$robust aggregator. For being precise in our interpretation, we recall below the Byz-DASHA-PAGE algorithm.

Algorithm 2: Byz-DASHA-PAGE [Rammal et al., 2024]

Input: Initial model $\theta^0 \in \mathbb{R}^d$ (chosen arbitrarily), initial update vector $R^0 = 0 \in \mathbb{R}^d$, total iterations $T \geq 1$, learning rate γ, robust aggregator F, momentum coefficient $\varrho \in [0,1)$ and, for each honest worker w_i: $g_i^0 = 0 \in \mathbb{R}^d$, $h_i^0 = 0 \in \mathbb{R}^d$, $m_i^0 = 0 \in \mathbb{R}^d$, and an unbiased compressor $\mathcal{Q}_i$ with compression parameter ω.

For $t = 1$ to T:

Server sets

$$c^t = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1-p \end{cases}.$$

Server broadcasts R^{t-1} to all the workers.

For each **honest worker** i do:

  Compute the updated model $\theta^t = \theta^{t-1} - \gamma R^t$.

  If $c^t = 1$ **then**

    Compute $h_i^t = \nabla \mathcal{L}_i\left(\theta^{t-1}\right)$.

  else

    Compute $h_i^t = h_i^{t-1} + \widehat{\Delta}(\theta^t, \theta^{t-1})$,

    where $\widehat{\Delta}(\theta^t, \theta^{t-1})$ is an unbiased stochastic estimate of $\Delta(\theta^t, \theta^{t-1}) = \nabla \mathcal{L}_i\left(\theta^t\right) - \nabla \mathcal{L}_i\left(\theta^{t-1}\right)$.

  Compute $m_i^t = \mathcal{Q}_i\left(h_i^t - h_i^{t-1} - \varrho(g_i^{t-1} - h_i^{t-1})\right)$.

  Compute $g_i^t = g_i^{t-1} + m_i^t$.

  Send m_i^t to the Server.

Server computes $R^t = F(m_1^t, \ldots, m_n^t)$.

Output: $\hat{\theta}_T$, chosen uniformly at random from $\{\theta^0, \ldots, \theta^{T-1}\}$.

C.1 From (f,κ)-robust aggregation to (δ, c)-robust aggregation

Recall from Section 2 that our work uses (f,κ)-robust aggregation, defined as: An aggregation rule $F : (\mathbb{R}^d)^n \to \mathbb{R}^d$ is said to be (f,κ)-robust if, for any set of n vectors $\{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ and any subset $S \subseteq [n]$ of size $n - f$, it holds that

$$\|F(x_1, \ldots, x_n) - \overline{x}_S\|^2 \leq \frac{\kappa}{|S|} \sum_{i \in S} \|x_i - \overline{x}_S\|^2,$$

where $\overline{x}_S \coloneqq \frac{1}{|S|} \sum_{i \in S} x_i$.

On the other hand, Byz-DASHA-PAGE [Rammal et al., 2024] uses a (δ, c)-RAgg, defined via pairwise dispersion on the honest set $\mathcal{H}$ (with $|\mathcal{H}| \geq (1-\delta)n$):

$$\frac{1}{|\mathcal{H}|(|\mathcal{H}|-1)} \sum_{i \neq \ell \in \mathcal{H}} \mathbb{E}\|x_i - x_\ell\|^2 \;\leq\; \sigma^2, \qquad \mathbb{E}\|\hat{x} - \bar{x}\|^2 \;\leq\; c\,\delta\,\sigma^2,$$

where $\bar{x} = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} x_i$.

Using the identity

$$\frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \|x_i - \bar{x}\|^2 = \frac{|\mathcal{H}| - 1}{2|\mathcal{H}|} \cdot \frac{1}{|\mathcal{H}|(|\mathcal{H}| - 1)} \sum_{i \neq \ell} \|x_i - x_\ell\|^2,$$

we obtain (f,κ)-robust averaging implies (δ, c)-RAgg with

$$c\,\delta \;=\; \tfrac{\kappa}{2}\left(1 - \tfrac{1}{|\mathcal{H}|}\right),$$

where $|\mathcal{H}| = n - f$ and $\delta = \frac{f}{n}$.

C.2 Notation mapping from [Rammal et al., 2024]

- Honest set size: $|\mathcal{H}| = n - f$ and $\delta = f/n$.
- Heterogeneity: Byz-DASHA-PAGE's B and ζ^2 correspond to our B^2 and G^2, respectively.
- Robustness constants: with $\tilde{\kappa} \coloneqq \kappa\left(1 - \frac{1}{n-f}\right)$, we have $8c\delta = 4\tilde{\kappa}$.
- Unbiased compression parameter ω: for RandK, $\omega = \frac{d}{k} - 1$. Writing $\alpha = d/k$, this is $\omega = \alpha - 1$.

C.3 Interpreting Theorem 2.2 in [Rammal et al., 2024] in our setup

Theorem 2.2 in [Rammal et al., 2024] states that for

$$\begin{aligned}
&\gamma \leq \frac{1}{L + \sqrt{\eta}}, \\
&\eta = \left(8\omega(2\omega+1)(L_\pm^2 + L^2) + \frac{1-p}{b}\left(12\omega(2\omega+1) + \tfrac{2}{p}\right)L_\pm^2\right)\left(\tfrac{1}{\sqrt{|H|}} + \sqrt{8c\delta}\right)^2, \\
&\delta < \left((8c + 4\sqrt{c})\, B\right)^{-1}
\end{aligned}$$

it holds

$$\mathbb{E}\|\nabla f(\hat{\theta})\|^2 \;\leq\; \frac{1}{A}\left(\frac{2\delta_0}{\gamma T} + \left(8c\delta + \sqrt{\tfrac{8c\delta}{|\mathcal{H}|}}\right)\zeta^2\right), \quad \text{with } A = 1 - \left(8c\delta + \sqrt{\tfrac{8c\delta}{|\mathcal{H}|}}\right)B.$$

Rewriting the theorem in our notation, we obtain that for:

$$\begin{aligned}
&\gamma \leq \frac{1}{L + \sqrt{\eta}}, \\
&\eta = \left(8\omega(2\omega+1)(L_\pm^2 + L^2) + \frac{1-p}{b}\left(12\omega(2\omega+1) + \tfrac{2}{p}\right)L_\pm^2\right)\left(\tfrac{1}{\sqrt{|H|}} + \sqrt{8c\delta}\right)^2, \\
&\left(4\tilde{\kappa} + 4\sqrt{\tfrac{\tilde{\kappa}\, f/n}{2}}\right) B^2 < 1,
\end{aligned}$$

it holds

$$\mathbb{E}\|\nabla L_H(\hat{\theta}_T)\|^2 \;\leq\; \frac{1}{1 - \left(4\tilde{\kappa} + 2\sqrt{\frac{\tilde{\kappa}}{n-f}}\right)B^2}\left(\frac{2\delta_0}{\gamma T} + \left(4\tilde{\kappa} + 2\sqrt{\tfrac{\tilde{\kappa}}{n-f}}\right)G^2\right).$$

For RoSDHB, the theoretical bounds hold for full gradient descent, hence $p = 1$. Moreover, we analyze our algorithm specifically for RandK unbiased compressor, hence $\omega = \frac{d}{k} - 1 = \alpha - 1$. Choosing the largest admissible stepsize $\gamma_{\max} = (L + \sqrt{\eta})^{-1}$ and considering the (favorable) case when $L_\pm = 0$, we obtain that for

$$\left(4\tilde{\kappa} + 4\sqrt{\tfrac{\tilde{\kappa}\, f/n}{2}}\right) B^2 < 1,$$

it holds that

$$\begin{aligned}
\mathbb{E}\|\nabla L_H(\hat{\theta}_T)\|^2 \;\leq\; &\frac{1}{1 - \left(4\tilde{\kappa} + 2\sqrt{\frac{\tilde{\kappa}}{n-f}}\right)B^2} \\
&\times \left[\frac{2L\,\delta_0}{T}\left(1 + \sqrt{8(\alpha-1)(2\alpha-1)}\left(\tfrac{1}{\sqrt{n-f}} + 2\sqrt{\tilde{\kappa}}\right)\right) + \left(4\tilde{\kappa} + 2\sqrt{\tfrac{\tilde{\kappa}}{n-f}}\right)G^2\right]. \qquad (24)
\end{aligned}$$

For $n - f \geq 2$, it holds that $\tilde{\kappa} \geq \kappa/2$. As our goal is to compare our algorithm to Byz-DASHA-PAGE in the most favorable conditions to Byz-DASHA-PAGE, we consider the less restrictive condition

$$\kappa B^2 < 1/2.$$

Using the above in the RHS of (24), we obtain the following lower bound:

$$\begin{aligned}\mathbb{E}\|\nabla L_H(\hat{\theta}_T)\|^2 &\leq \frac{1}{1-2\kappa B^2}\left[\frac{2L\,\delta_0}{T}\left(1+4(\alpha-1)\left(\frac{1}{\sqrt{2n}}+\sqrt{\kappa}\right)\right)+\frac{\kappa}{2}G^2\right] \\ &\leq \mathcal{O}\left(\frac{1+4(\alpha-1)\left(\frac{1}{\sqrt{n}}+\sqrt{\kappa}\right)}{(1-2\kappa B^2)T}+\frac{\kappa G^2}{1-2\kappa B^2}\right).\end{aligned}$$

D Additional Empirical Details & Results

D.1 Byzantine attacks and defense

To evaluate the resilience of our algorithm, we simulate two Byzantine attack strategies: *Fall of Empires (FOE)* [Xie et al., 2020] and *A Little Is Enough (ALIE)* [Baruch et al., 2019]. Both of these attacks rely on crafting malicious updates that bias the aggregated updates while aiming to remaining undetected by robust defenses. We briefly describe them below.

Fix a time step t. Let B^t denote the vectors sent by the byzantine workers in time step t. Recall from Section 2 that the byzantine workers can collude seamlessly, know the learning algorithm used by the server, and observe all messages exchanged between the server and the honest workers throughout training. For $i \in \mathcal{H}$, suppose m_i^t is the momentum of the honest worker w_i, and define $\eta \geq 0$ to be a fixed constant. Then, we have:

Fall of Empires (FOE). The byzantine updates are set as:

$$B^t = (1-\eta)\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} m_i^t.$$

A Little Is Enough (ALIE). Let $\sigma_{\mathcal{H}}^t$ be the coordinate-wise standard deviation of the momentum vectors recieved by the server from honest workers in time step t. Then, the byzantine updates are set as:

$$B^t = \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} m_i^t + \eta\sigma_{\mathcal{H}}^t$$

In our experiments for RoSDHB, we determine optimal value of η that maximizes the L2-distance between the aggregated update R^t and mean of the honest momentum vectors received by the server. Note that the index of coordinates sent by the byzantine workers are the same as those of the honest workers for our algorithm, since they need to adhere to the mask sent by the server in time step t. For Byz-DASHA-PAGE, since the server does not decide the mask in every time step, workers have the flexibility to determine their mask independently. We incorporate this in the attack design for Byz-DASHA-PAGE in the following manner: The byzantine workers independently sample the RandK mask in each time step, and modify only those masked coordinates for the attack strategy.

Robust Aggregation. In federated learning, the server aggregates the received updates to compute the global model in an iteration. In the presence of adversarial workers, the goal of the aggregator is to compute a function over the gradients such that the distance between the resulting global gradient and the gradient of the honest workers is minimized. We assume that F is an (f,κ)-robust aggregator. In our experiments, we employ the Coordinate-Wise Trimmed Mean (CWTM) robust aggregation method from [Yin et al., 2018].

Definition D.1 (Coordinate-Wise-Trimmed Mean)**.** *Fix a time step. Suppose $g_1, ..., g_n \in \mathbb{R}^d$ are the gradient updates of the workers in the system, and f is the number of byzantine workers in the system. For each coordinate $k \in [d]$, let $[\overline{g}_1]_k \leq [\overline{g}_2]_k \leq \cdots \leq [\overline{g}_n]_k$ be the sorted values of k^{th} coordinate of the n gradient updates. Then, the coordinate-wise trimmed mean of $g_1, \ldots, g_n$, denoted by $CWTM(g_1, \ldots g_n)$, is a vector in $\mathbb{R}$ such that the k^{th} coordinate is given by:*

$$[CWTM(g_1, \ldots g_n)]_k = \frac{1}{n-2f}\sum_{i=f+1}^{n-f} [\overline{g}_i]_k.$$

D.2 Experimental setup details

System specification All experiments were conducted on a server equipped with four NVIDIA L40S GPUs (45 GiB each), two AMD EPYC 9124 16-core processors, and 512 GiB of RAM. The software environment consisted of Python 3.10.6 and PyTorch 2.6.0, running with CUDA 12.4 and cuDNN 9.1 on Debian GNU/Linux 11 (Bullseye).

Hyperparameters

Compression	MNIST					CIFAR-10			
Ratio, k/d	0.05	0.10	0.30	0.50	1.00	0.25	0.5	0.75	1.00
Learning Rate γ	0.16	0.16	0.4	0.4	0.4	0.7	0.7	0.7	0.7
Grid Search for γ	{0.008,0.04,0.08, 0.16, 0.4}					{0.1,0.2,0.3,0.5,0.7}			
Momentum, β	0.8					0.8			
Batch Size	full batch 6000					128			
# Byzantine Workers	$f = \{1, 3, 5\}$					$f = \{1, 3, 5\}$			

Table 1: Summary of hyperparameters used for RoSDHB experiments when data partition follows a symmetric Dirichlet distribution with parameters $w = 5$.

Compression	MNIST					CIFAR-10			
Ratio, k/d	0.05	0.10	0.30	0.50	1.00	0.25	0.5	0.75	1.00
Learning Rate γ	0.008	0.008	0.008	0.08	0.16	0.03	0.03	0.03	0.03
Grid Search for γ	{0.004, 0.0056, 0.008, 0.04, 0.08, 0.16, 0.4}					{ 0.01,0.03,0.05, 0.1,0.2}			
Momentum, ϱ	$1/(2(d/k) - 1)$					$1/(2(d/k) - 1)$			
Batch Size	full batch 6000					128			
# Byzantine Workers	$f = \{1, 3, 5\}$					$f = \{1, 3, 5\}$			

Table 2: Summary of hyper-parameters used for Byz-DASHA-PAGE (described in Algorithm 2) experiments when data partition follows a symmetric Dirichlet distribution with parameters $w = 5$.[2]

Heterogeneity	RoSDHB				Byz-DASHA-PAGE			
parameter, w	0.5	1.0	3.0	5.0	0.5	1.0	3.0	5.0
Learning Rate γ	0.4	0.4	0.4	0.16	0.04	0.008	0.008	0.008
Grid Search for γ	{0.008,0.04,0.08, 0.16, 0.4}				{0.008,0.04,0.08, 0.16, 0.4}			
Momentum	$\beta = 0.8$				$\varrho = 1/19$			

Table: Summary of hyperparameters used for RoSDHB and Byz-DASHA-PAGE experiments on MNIST dataset with $k/d = 0.1$, $f = 1$ and full batch under different data partitions follows a symmetric Dirichlet distribution with parameters w (Experiment in Figure 4).

[2]We chose the optimal momentum α for Byz-DASHA-PAGE as suggested in their paper.

D.3 Further interpretations of the main results

Accuracy	RoSDHB (f=1)	SOTA (f=1)	Speedup (f=1)	RoSDHB (f=3)	SOTA (f=3)	Speedup (f=3)
0.76	10.8 ± 1.04	55.8 ± 7.87	**5.17**×	17.0 ± 1.17	93.4 ± 11.93	**5.49**×
0.78	12.0 ± 1.02	61.8 ± 8.32	**5.15**×	19.0 ± 1.17	104.4 ± 13.11	**5.49**×
0.80	13.0 ± 1.23	70.2 ± 8.89	**5.40**×	22.8 ± 1.18	117.0 ± 14.60	**5.13**×
0.82	15.6 ± 1.37	82.8 ± 9.97	**5.31**×	25.6 ± 1.25	133.6 ± 16.52	**5.22**×
0.84	18.8 ± 1.15	102.6 ± 10.93	**5.46**×	31.2 ± 1.91	162.2 ± 19.79	**5.20**×
0.86	23.0 ± 1.10	135.8 ± 12.20	**5.90**×	37.6 ± 2.54	168.0 ± 10.03	**4.47**×
0.88	32.0 ± 1.36	201.3 ± 7.49	**6.29**×	53.0 ± 3.09	–	–
0.90	56.8 ± 2.60	–	–	90.8 ± 5.06	–	–

Table 3: Convergence time (mean ± std) required to reach various test threshold accuracy on MNIST with compression ratio $k/d = 0.1$, comparing RoSDHB and Byz-DASHA-PAGE (SOTA) under $f = 1, 3$ Byzantine workers . RoSDHB consistently outperforms SOTA, achieving up to 6.3× speedup.

Accuracy	RoSDHB (f=1)	SOTA (f=1)	Speedup (f=1)	RoSDHB (f=3)	SOTA (f=3)	Speedup (f=3)
0.30	13.2 ± 0.36	54.4 ± 2.11	**4.12**×	26.4 ± 0.18	87.6 ± 3.17	**3.32**×
0.32	15.6 ± 0.33	66.0 ± 2.42	**4.23**×	28.8 ± 0.78	98.0 ± 3.66	**3.40**×
0.34	17.2 ± 0.22	78.0 ± 2.37	**4.53**×	36.8 ± 0.61	108.0 ± 3.58	**2.93**×
0.36	19.6 ± 0.44	90.0 ± 3.16	**4.59**×	43.6 ± 1.21	125.6 ± 3.00	**2.88**×
0.38	23.6 ± 0.33	106.4 ± 3.65	**4.51**×	55.2 ± 1.99	151.6 ± 4.28	**2.75**×
0.40	26.0 ± 0.69	139.2 ± 5.79	**5.35**×	80.4 ± 3.12	182.8 ± 2.48	**2.27**×
0.42	29.6 ± 0.59	164.8 ± 5.70	**5.57**×	98.4 ± 2.30	248.0 ± 6.34	**2.52**×
0.44	33.6 ± 0.44	210.5 ± 8.71	**6.26**×	111.6 ± 3.18	–	–

Table 4: Convergence time (mean ± std in epochs) to reach various test threshold accuracy on CIFAR-10 for compression ratio $k/d = 0.5$, comparing RoSDHB and Byz-DASHA-PAGE (SOTA) under different numbers of Byzantine workers $f = 1$ and $f = 3$.

f	k/d	SOTA Time		RoSDHB Time		Speedup	SOTA Coords		RoSDHB Coords		Comm.
		Mean	Std	Mean	Std	$\frac{\text{SOTA}}{\text{RoSDHB}}$	Mean	Std	Mean	Std	Savings (%)
1	0.05	161.0	21.81	81.0	8.84	1.99×	95231.5	12902.84	47911.5	5229.34	49.67
1	0.1	117.0	11.44	20.6	1.28	5.68×	138411.0	13537.99	24369.8	1518.67	82.38
1	0.3	109.2	11.65	12.0	0.40	9.10×	387550.8	41332.06	42588.0	1419.60	89.01
1	0.5	95.25	30.41	10.8	0.72	8.82×	563403.8	179865.79	63882.0	4232.43	88.65
1	1.0	19.6	2.63	10.6	0.73	1.85×	231868.0	31137.85	125398.0	8596.10	45.93
3	0.05	193.0	14.35	133.8	14.91	1.44×	114159.5	8486.32	79142.7	8820.14	30.64
3	0.1	185.4	23.34	33.4	1.54	5.55×	219328.2	27609.51	39512.2	1820.44	82.00
3	0.3	191.5	13.81	21.0	1.52	9.12×	679633.5	49024.08	74529.0	5405.68	89.01
3	0.5	117.0	34.34	18.4	1.43	6.36×	692055.0	203148.65	108836.0	8464.86	84.28
3	1.0	34.8	4.88	18.2	1.11	1.91×	411684.0	57703.24	215306.0	13130.77	47.70

Table 5: Convergence time (mean ± std) and number of coordinates communicated (mean ± std) to achieve a threshold accuracy of 85% for Byz-DASHA-PAGE (SOTA) and RoSDHB on MNIST dataset under various compression ratios (k/d) and Byzantine settings (f). Speedup indicates how many times faster RoSDHB converges compared to SOTA. Communication savings shows percentage reduction in communicated coordinates by RoSDHB compared to SOTA computed as Communication Savings (%) $= \left(\frac{\text{SOTA Mean} - \text{RoSDHB Mean}}{\text{SOTA Mean}}\right) \times 100$.

f	k/d	SOTA Time		RoSDHB Time		Speedup	SOTA Coords		RoSDHB Coords		Comm.
		Mean	Std	Mean	Std	$\frac{\text{SOTA}}{\text{RoSDHB}}$	Mean	Std	Mean	Std	Savings (%)
1	0.25	127.2	7.28	25.2	0.44	5.05×	382.6M	21.90M	70.4M	1.22M	81.59
1	0.5	139.2	11.57	26.0	1.39	5.35×	797.6M	66.32M	145.3M	7.74M	81.79
1	0.75	150.0	8.52	28.4	1.31	5.28×	1267.8M	72.04M	238.0M	11.02M	81.22
3	0.25	151.6	10.90	86.8	3.47	1.75×	456.0M	32.79M	242.5M	9.69M	46.84
3	0.5	182.8	4.95	80.4	6.23	2.27×	1047.4M	28.37M	449.2M	34.82M	57.12
3	0.75	246.0	11.11	67.6	4.94	3.64×	2079.2M	93.93M	566.5M	41.44M	72.76

Table 6: Convergence time (mean ± std) and number of coordinates communicated (mean ± std) to achieve a threshold accuracy of 40% for Byz-DASHA-PAGE (SOTA) and RoSDHB on CIFAR-10 under different compression ratios (k/d) and Byzantine worker settings (f).

D.4 Full set of experiments

We present our experimental results where the system is under the FOE attack and the ALIE attack for the MNIST and CIFAR-10 datasets.

We model a system consisting of 10 honest workers. We consider three Byzantine regimes: $f = 1$, $f = 3$, and $f = 5$. We evaluate the performance of our algorithm under the FOE and ALIE attacks and compare against SOTA approach - Byz-DASHA-PAGE from [Rammal et al., 2024] for the CWTM aggregation rule (detailed in Appendix D.1). The plots are presented below and complement our (partial) results in Section 4 of the main paper.

Comprehensive results for FOE attack

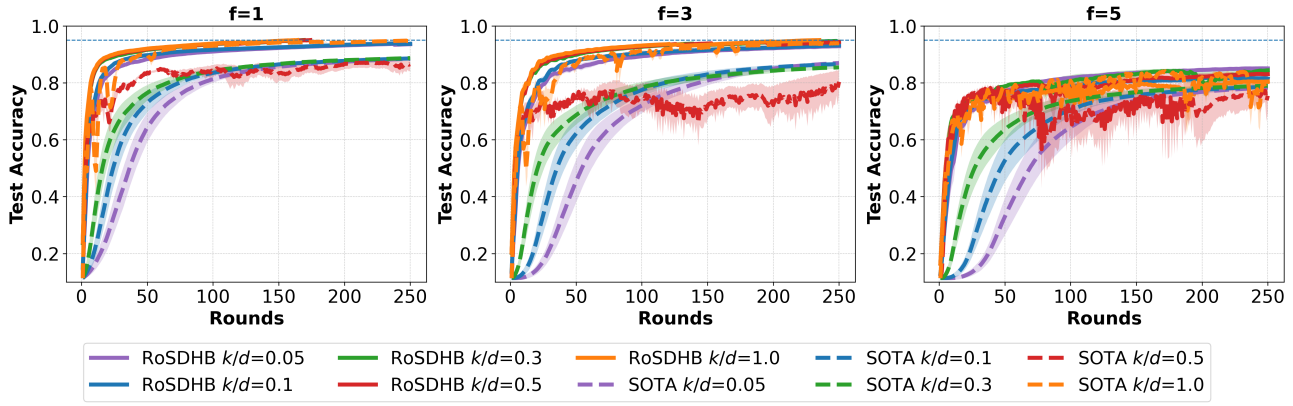


Figure 5: The convergence plots of RoSDHB and Byz-DASHA-PAGE (SOTA) on MNIST under the FOE attack for varying sampling ratio $k/d \in \{0.05, 0.1, 0.3, 0.5, 1.0\}$ and number of byzantine workers $f \in \{1, 3, 5\}$.

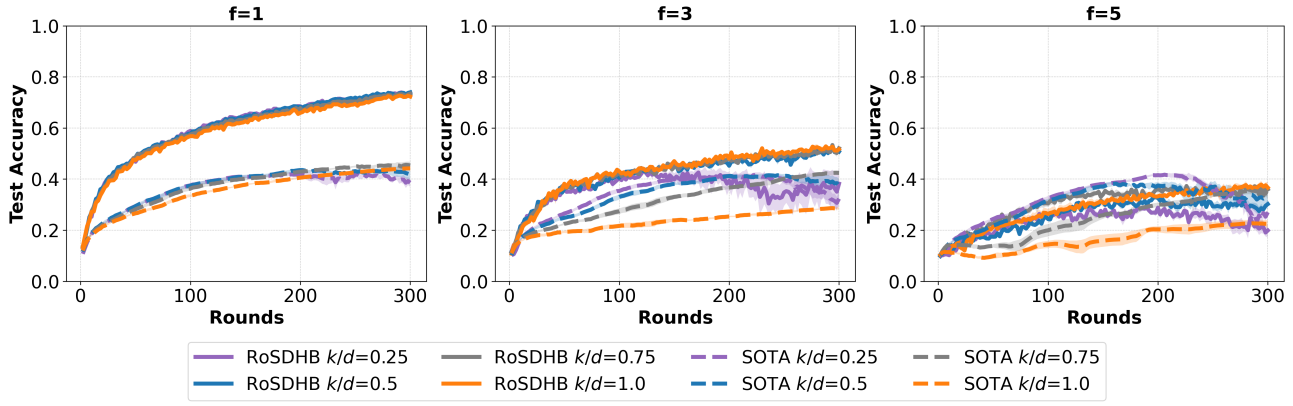


Figure 6: The convergence plots of stochastic version of RoSDHB and Byz-DASHA-PAGE (SOTA) on CIFAR-10 under the FOE attack for varying sampling ratio $k/d \in \{0.25, 0.5, 0.75, 1.0\}$ and number of byzantine workers $f \in \{1, 3, 5\}$.

Comprehensive results for ALIE attack

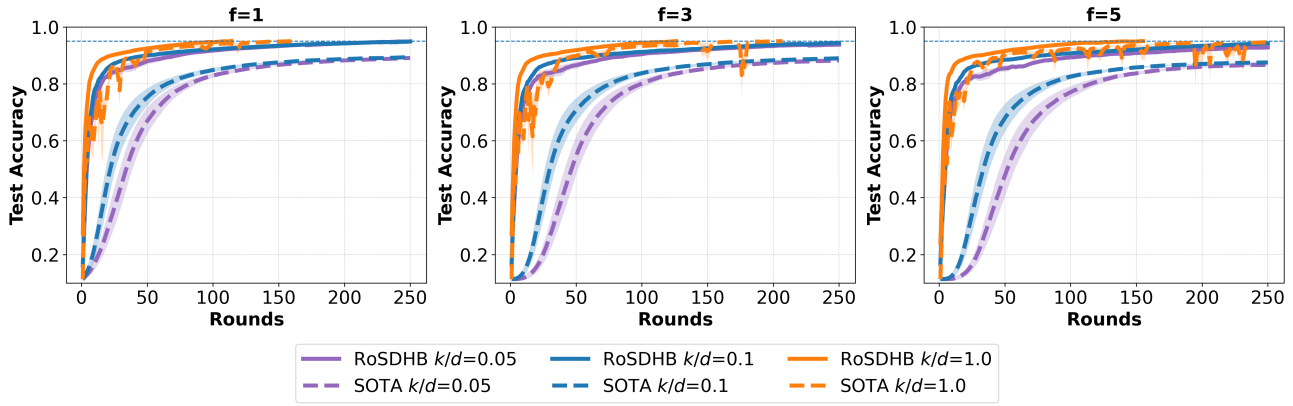


Figure 7: The convergence plots of RoSDHB and Byz-DASHA-PAGE (SOTA) on MNIST under the ALIE attack for varying sampling ratio $k/d \in \{0.05, 0.1, 1.0\}$ and number of byzantine workers $f \in \{1, 3, 5\}$.

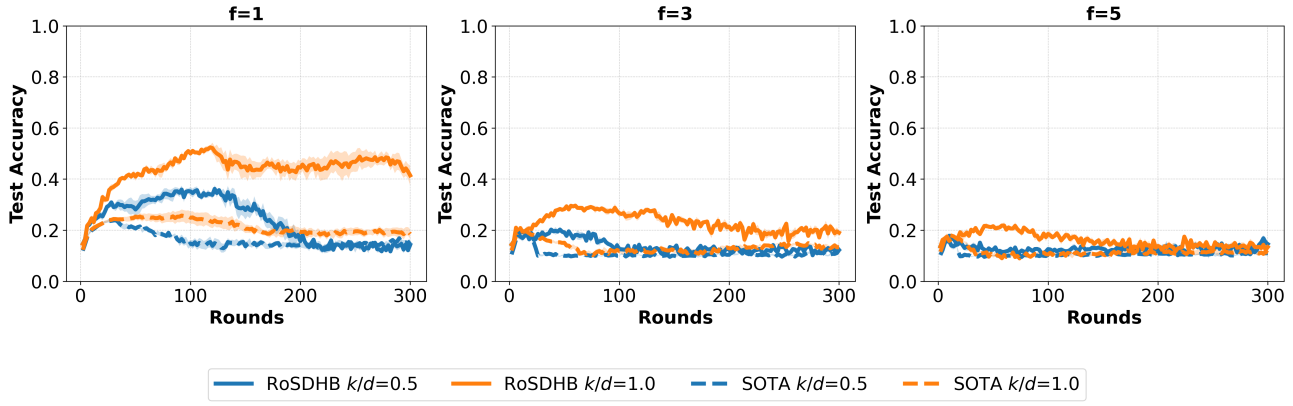


Figure 8: The convergence plots of stochastic version of RoSDHB and Byz-DASHA-PAGE (SOTA) on CIFAR-10 under the ALIE attack for varying sampling ratio $k/d \in \{0.5, 1.0\}$ and number of byzantine workers $f \in \{1, 3, 5\}$.

D.5 RoSDHB vs Byz-EF21

Table 7: Rounds to reach 0.90 test accuracy on MNIST dataset and corresponding speedups of RoSDHB over Byz-DASHA-PAGE (SOTA) and Byz-EF21 with Top-k compressor. Speedup is computed as (baseline / RoSDHB). Byz-DASHA-PAGE fails to reach the target accuracy under high compression ratios.

	$f = 1$					$f = 3$				
k/d	**RoSDHB**	**SOTA**	**Speedup**	**Byz-EF21**	**Speedup**	**RoSDHB**	**SOTA**	**Speedup**	**Byz-EF21**	**Speedup**
0.05	76.8 ± 8.20	–	–	84.8 ± 14.82	**1.10**×	119.0 ± 11.38	–	–	168.6 ± 32.14	**1.42**×
0.10	56.8 ± 2.59	–	–	65.8 ± 5.82	**1.16**×	90.8 ± 5.06	–	–	91.2 ± 14.23	**1.00**×
1.00	$\mathbf{25.0 \pm 0}$	45.0 ± 3.43	**1.80**×	45.6 ± 3.40	**1.82**×	$\mathbf{37.4 \pm 1.28}$	61.6 ± 5.43	**1.65**×	65.8 ± 5.31	**1.76**×

From Table 7, Byz-EF21 with the Top-k compressor outperforms Byz-DASHA-PAGE but remains inferior to RoSDHB, which achieves up to a 1.82× speedup. Note that Byz-EF21 only enjoys a suboptimal theoretical convergence guarantee.