Dynamic Byzantine-Robust Learning: Adapting to Switching Byzantine Workers

Ron Dorfman [1] **Naseem Yehya** [1] **Kfir Y. Levy** [1]

Abstract

Byzantine-robust learning has emerged as a prominent fault-tolerant distributed machine learning framework. However, most techniques focus on the *static* setting, wherein the identity of Byzantine workers remains unchanged throughout the learning process. This assumption fails to capture real-world *dynamic* Byzantine behaviors, which may include intermittent malfunctions or targeted, time-limited attacks. Addressing this limitation, we propose DynaBRO – a new method capable of withstanding any sublinear number of identity changes across rounds. Specifically, when the number of such changes is $\mathcal{O}(\sqrt{T})$ (where T is the total number of training rounds), DynaBRO nearly matches the state-of-the-art asymptotic convergence rate of the static setting. Our method utilizes a multi-level Monte Carlo (MLMC) gradient estimation technique applied at the server to robustly aggregated worker updates. By additionally leveraging an adaptive learning rate, we circumvent the need for prior knowledge of the fraction of Byzantine workers.

1. Introduction

Recently, there has been an increasing interest in large-scale distributed machine learning (ML), where multiple machines (i.e., *workers*) collaboratively aim at minimizing some global objective under the coordination of a central *server* (Verbraeken et al., 2020; Kairouz et al., 2021). This approach, leveraging the collective power of multiple computation nodes, holds the promise of significantly reducing training times for complex ML models, such as large language models (Brown et al., 2020). However, the growing reliance on distributed ML systems exposes them to potential errors, malfunctions, and even adversarial attacks. These *Byzantine* faults pose a significant risk to the integrity of the learning process and could lead to reduced reliability and accuracy in the predictions of the resulting models (Lamport et al., 2019; Guerraoui et al., 2023).

Due to its significance in distributed ML, Byzantine fault-tolerance has attracted considerable interest. Indeed, many prior works have focused on Byzantine-robust learning, aiming to ensure effective learning in the presence of Byzantine machines. These efforts have led to the development of algorithms capable of enduring Byzantine attacks (Feng et al., 2014; Su & Vaidya, 2016; Blanchard et al., 2017; Chen et al., 2017; Alistarh et al., 2018; Guerraoui et al., 2018; Yin et al., 2018; Allen-Zhu et al., 2020; Karimireddy et al., 2021; 2022; Farhadkhani et al., 2022; Allouah et al., 2023).

While the existing body of research has significantly advanced the understanding of Byzantine-robust learning, a notable gap persists in the treatment of the problem. The vast majority of research in this area has focused on the static setting, wherein the identity of Byzantine workers remains fixed throughout the entire learning process. Nevertheless, real-world distributed learning systems may often encounter *dynamic* Byzantine behaviors, where machines exhibit erratic and unpredictable fault patterns. Despite the importance of these scenarios, the investigation of robustness against such dynamic behaviors remains underexplored.

In federated learning, for instance, the concept of partial participation inherently introduces a dynamic aspect (Bonawitz et al., 2019; Kairouz et al., 2021). Workers typically join and leave the training process, leading to a scenario where a Byzantine worker might be present in one round and absent in the next. In fact, the same node might switch between honest and Byzantine behaviors; such fluctuations could stem from strategic manipulations by an adversary seeking to evade detection, or due to software updates that inadvertently trigger or resolve certain security vulnerabilities. Another domain includes volunteer computing paradigms (Kijsipongse et al., 2018; Ryabinin & Gusev, 2020; Atre et al., 2021), characterized by a large pool of less reliable workers and regular occurrences of node failures. Prolonged training times of complex ML models in these settings often result in intermittent node failures, typically due to hardware issues, network instabilities, or maintenance activities. These domains, where dynamic Byzantine behaviors are prevalent, pose challenges unaddressed by the static approach.

[1]Department of Electrical and Computer Engineering, Technion, Haifa, Israel. Correspondence to: Ron Dorfman <rdorfman@campus.technion.ac.il>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Previous research has established that convergence is unattainable when Byzantine workers change their identities in each round (Karimireddy et al., 2021).[1] However, it remains unclear how a **limited** number of identity-switching rounds affects the convergence rate. This work addresses this challenge by introducing a new method we term DynaBRO. We establish its convergence, noting that it maintains the asymptotic convergence rate of the static setting as long as the number of rounds featuring Byzantine behavior alterations does not exceed $\mathcal{O}(\sqrt{T})$, where T is the total number of training rounds. Beyond this threshold, the convergence rates begin to degrade linearly with the increase in the number of such rounds.

The key ingredient of our approach is a multi-level Monte Carlo (MLMC) gradient estimation technique (Dorfman & Levy, 2022; Beznosikov et al., 2023), serving as a bias-reduction tool. In Section 3, we show that combining it with a large class of aggregation rules (Allouah et al., 2023) mitigates the bias introduced by Byzantine workers in the static setting, an analysis of independent interest. Transitioning to the dynamic setting in Section 4, we refine the MLMC estimator with an added fail-safe filter to address its inherent susceptibility to dynamic Byzantine changes that could introduce a significant bias. This modification is crucial due to the estimator's reliance on multiple consecutive samples. Then, in Section 5, we shift our focus to optimality and adaptivity. The introduction of a new aggregation rule enables us to derive asymptotically optimal convergence bounds for a limited number of identity-switching rounds. Furthermore, by employing an adaptive learning rate, we eliminate the need for prior knowledge of the objective's smoothness and the fraction of Byzantine workers. Finally, in Section 6, we explore the practical aspects and benefits of our approach through experiments on image classification tasks with two dynamic identity-switching strategies.

2. Preliminaries and Related Work

In this section, we formalize our objective, specify our assumptions, and overview relevant related work.

2.1. Problem Formulation and Assumptions

We consider stochastic optimization problems, where the objective is to minimize the expected loss given an unknown data distribution $\mathcal{D}$ and a set of loss functions $\{x \mapsto F(\cdot;\xi)\}$ parameterized by $\xi \sim \mathcal{D}$. Formally, our goal is to solve:

$$\min_{x\in\mathcal{K}} f(x) \coloneqq \mathbb{E}_{\xi\sim\mathcal{D}}\left[F(x;\xi)\right] , \tag{1}$$

where $\mathcal{K} \subseteq \mathbb{R}^d$ is the optimization domain. To this end, we assume there are m workers (i.e., computations nodes), each with access to samples from $\mathcal{D}$. This homogeneous setting was previously studied in the context of Byzantine-robust learning (Blanchard et al., 2017; Alistarh et al., 2018; Allen-Zhu et al., 2020), and it is realistic in collaborative learning scenarios, where workers may have access to the entire dataset (Kijsipongse et al., 2018; Diskin et al., 2021; Gorbunov et al., 2022). For each round t, we define $\mathcal{G}_t \subseteq \{1,\dots,m\}$ as the set of honest workers, adhering to the prescribed protocol; the remaining workers are Byzantine and may send arbitrary vectors to the server. Notably, in the static setting, the identity of honest workers is fixed over time, i.e., $\mathcal{G}_t$ is identical across rounds. Allen-Zhu et al. (2020) refer to this dynamic model as *ID relabeling*. For simplicity, we assume the fraction of Byzantine workers is fixed across rounds and denote it by $\delta \coloneqq 1-|\mathcal{G}_1|/m < 1/2$.

We focus on smooth minimization, namely, we assume the objective f is L-smooth, i.e., for every $x, y \in \mathcal{K}$, it holds that $f(y) \le f(x) + \nabla f(x)^\top (y-x) + \frac{L}{2}\|y-x\|^2$. Additionally, we will adopt one of the following two assumptions regarding the stochastic gradient noise.

Assumption 2.1 (Bounded variance)**.** For every $x \in \mathcal{K}$,

$$\mathbb{E}_{\xi\sim\mathcal{D}}\|\nabla F(x;\xi) - \nabla f(x)\|^2 \le \sigma^2 .$$

This standard assumption in Byzantine-robust optimization (Karimireddy et al., 2021; 2022; Farhadkhani et al., 2022; Allouah et al., 2023) is utilized in Section 3 to establish intuition in the static setting. For the dynamic case, we require a stronger assumption of deterministically bounded noise (Alistarh et al., 2018; Allen-Zhu et al., 2020).[2]

Assumption 2.2 (Bounded noise)**.** For every $x \in \mathcal{K}$, $\xi \sim \mathcal{D}$,

$$\|\nabla F(x;\xi) - \nabla f(x)\| \le \mathcal{V} .$$

We provide convergence bounds for both convex and non-convex problems, presenting only non-convex results for brevity and moving convex analyses to the appendix.[3] For the convex case, we assume a bounded domain, as follows:

Assumption 2.3 (Bounded domain)**.** The domain $\mathcal{K}$ is convex, compact, and for every $x, y \in \mathcal{K}$: $\|x-y\| \le D$.

Notations. Throughout, $\|\cdot\|$ represents the L_2 norm, and for any $n \in \mathbb{N}$, $[n] \coloneqq \{1,\dots,n\}$. We denote any optimal solution of (1) by x^* and its corresponding optimal value by f^*. Additionally, we define $\mathcal{F}_t$ as the filtration at round t, encompassing all prior randomness. Expectation and probability are denoted by $\mathbb{E}$ and $\mathbb{P}$, respectively, with $\mathbb{E}_t$ and $\mathbb{P}_t$ indicating their conditional counterparts given $\mathcal{F}_t$. Finally, we use standard big-O notation, where $\mathcal{O}(\cdot)$ hides numerical constants and $\tilde{\mathcal{O}}(\cdot)$ omits poly-logarithmic terms.

[1] The lower bound of Karimireddy et al. (2021) applies when the number of Byzantines is logarithmic in the number of rounds.

[2] While our analysis could extend to sub-Gaussian noise (up to logarithmic factors), it becomes more technical. Thus, following Allen-Zhu et al. (2020), we opt for bounded noise to ensure simplicity and technical clarity (see footnote 3 therein).

[3] Since global non-convex minimization is generally intractable, we focus on finding an approximate stationary point.

Table 1. Comparison of history-dependence in different Byzantine-robust techniques.

Method	Per-worker Cost	Window Size	Window Type
ByzantineSGD (Alistarh et al., 2018)	T	T	Deterministic
SafeguardSGD (Allen-Zhu et al., 2020)	T	$\mathcal{O}(T^{5/8})$	Deterministic
Worker-momentum (e.g., Karimireddy et al., 2021)	T	$\tilde{\mathcal{O}}(\sqrt{T})^*$	Deterministic
MLMC (ours)	$\mathcal{O}(T \log T)^\dagger$	$\mathcal{O}(\log T)^\dagger$	Stochastic

* For momentum parameter $\alpha := 1 - \beta \propto 1/\sqrt{T}$.
† In expectation.

2.2. Related Work

The importance of history for Byzantine-robustness. It has been well-established that history is critical for achieving Byzantine-robustness in the static setting. Karimireddy et al. (2021) were the first to formally prove that any memoryless method—where the update for each round depends only on computations performed in that round—may fail to converge. As we detail in Section 3.1, the importance of history arises from the ability of Byzantine workers to inject in each round bias proportional to the natural noise of honest workers, which is sufficient to circumvent convergence. Thus, employing a variance reduction technique, requiring some historical dependence, is crucial to mitigate the injected bias. Consequently, they proposed applying a robust aggregation rule to worker momentums instead of directly to gradients. By setting the momentum parameter as $\alpha := 1 - \beta \propto 1/\sqrt{T}$, they achieved state-of-the-art convergence guarantees in the presence of Byzantine workers. This momentum method has emerged as the leading approach for Byzantine-robust learning, with its efficacy also demonstrated through empirical evidence (El-Mhamdi et al., 2020; Farhadkhani et al., 2022; Allouah et al., 2023).

Additional history-dependent methods include ByzantineSGD (Alistarh et al., 2018) for convex minimization and SafeguardSGD (Allen-Zhu et al., 2020) for finding local minima of non-convex functions. Both techniques estimate the set of honest workers by tracking worker statistics (e.g., gradient-iterate products) and applying some median-based filtering. While ByzantineSGD relies on the entire history, SafeguardSGD uses information within windows of $T_1 \in \mathcal{O}(T^{5/8})$ rounds and incorporates a reset mechanism. Consequently, SafeguardSGD can withstand Byzantine identity changes occurring between these windows. However, there is no reason to assume that ID relabeling occurs only at specific rounds, and our method can handle such changes at any round, without restrictions on when the changes happen.

Although history dependence is crucial in static environments, it poses significant challenges in dynamic settings, in which history may become unreliable. For example, the computations of a Byzantine worker turning honest may still be influenced by prior misbehaviors if the history dependence window encompasses those rounds. Therefore, methods that rely on long historical information are vulnerable to such identity changes. Our approach relies on MLMC gradient estimation technique, which confines the history window size to $\mathcal{O}(\log T)$ in expectation; refer to Table 1 for a comparison between different history-dependent methods.

Byzantine-robustness and worker sampling. To date, we are only aware of two works that address the challenges of the dynamic setting (Data & Diggavi, 2021; Malinovsky et al., 2023), where the dynamic behavior stems from worker sampling, i.e., different workers may actively participate in each training round. Data & Diggavi (2021) were the first to study this challenging setting, providing convergence results for both strongly- and non-convex objectives. Yet, their upper bounds include a non-vanishing term proportional to the gradient noise, which is sub-optimal in the homogeneous setting. Malinovsky et al. (2023) improved upon previous limitations, proposing Byz-VR-MARINA-PP, which can handle some rounds dominated by Byzantine workers. Their work follows a rich body of literature on Byzantine-robust finite-sum minimization (Wu et al., 2020; Zhu & Ling, 2021; Gorbunov et al., 2022). Nonetheless, these studies do not provide excess loss (i.e., generalization) guarantees.

MLMC estimation. Originally utilized in the context of stochastic differential equations (Giles, 2015), MLMC methods have since been applied in various ML and optimization contexts. These include, for example, distributionally robust optimization (Levy et al., 2020; Hu et al., 2021) and latent variable models (Shi & Cornish, 2021). Asi et al. (2021) employed an MLMC optimum estimator for calculating proximal points and gradients of the Moreau-Yoshida envelope. Specifically for gradient estimation, MLMC is useful for efficiently generating low-bias gradients in scenarios where obtaining unbiased gradients is either impractical or computationally intensive, e.g., conditional stochastic optimization (Hu et al., 2021), stochastic optimization with Markovian noise (Dorfman & Levy, 2022; Beznosikov et al., 2023), and reinforcement learning (Suttle et al., 2023).

3. Warm-up: Static Robustness with MLMC

In this section, we study the static setting where Byzantine machine identities are fixed across time. We develop intuition in Section 3.1, illustrating how Byzantine machines can hinder convergence. In Section 3.2, we introduce our MLMC estimator, present its bias-variance properties, and highlight its role as a bias reduction technique. Finally, in Section 3.3, we show how integrating the MLMC estimator with a robust aggregation rule implies Byzantine-robustness.

3.1. Motivation

We examine the distributed stochastic gradient descent update rule, defined as $x_{t+1} = x_t - \eta\mathcal{A}(g_{t,1}, \dots, g_{t,m})$, with $\eta > 0$ as learning rate, $\mathcal{A} : \mathbb{R}^{d\times m} \to \mathbb{R}^d$ as an aggregation rule, and $g_{t,i}$ representing the message from worker $i \in [m]$ at time $t \in [T]$. In the Byzantine-free case, $\mathcal{A}$ typically averages the inputs, yet a single Byzantine worker can arbitrarily manipulate this aggregation result (Blanchard et al., 2017).

Ideally, $\mathcal{A}$ would isolate the Byzantine inputs and average the honest inputs, yielding a conditionally unbiased gradient estimate with reduced variance. However, Byzantine workers can blend in by aligning their messages closely with the expected noise range of honest gradients. Thus, they can inject a bias that is indistinguishable from the natural noise inherent to honest messages, thereby hindering convergence. Specifically, if honest messages deviate by σ from the true gradient, Byzantine workers can introduce a bias of $\mathcal{O}(\sigma\sqrt{\delta})$ at each round, effectively bounding convergence to a similarly scaled neighborhood (Ajalloeian & Stich, 2020).

Addressing this challenge, a straightforward mitigation strategy might involve all honest workers computing stochastic gradients across large mini-batches. This approach reduces their variance, thereby shrinking the feasible 'hiding region' for Byzantine messages. As we establish in Appendix A, theory suggests a mini-batch size of $\Omega(T)$ is required to ensure sufficiently low bias. However, this approach proves to be extremely inefficient, necessitating an excessive total of $\Omega(T^2)$ stochastic gradient evaluations per worker.

Instead of implicitly reducing Byzantine bias through honest worker variance reduction, we propose a direct bias reduction strategy by employing an MLMC gradient estimation technique post-aggregation, i.e., to the robustly aggregated gradients. In the next section, we introduce the MLMC estimator and establish its favorable bias-variance properties.

3.2. MLMC Gradient Estimation

Our MLMC estimator utilizes any mapping $\mathcal{M}_f : \mathbb{R}^d\times\mathbb{N} \to \mathbb{R}^d$ that, given a query vector x and a budget N (in terms of stochastic gradient evaluations), produces a vector whose mean squared error (MSE) in estimating $\nabla f(x)$ is inversely proportional to N. Formally, for some $c > 0$, we have

$$\mathbb{E}\left\|\mathcal{M}_f(x, N) - \nabla f(x)\right\|^2 \leq \frac{c^2}{N}, \quad \forall x \in \mathbb{R}^d, N \in \mathbb{N}\,. \quad (2)$$

The MLMC gradient estimator is defined as follows: Sample $J \sim \text{Geom}(1/2)$ and, denoting $g^j := \mathcal{M}_f(x, 2^j)$, set

$$g^{\text{MLMC}} = g^0 + \begin{cases} 2^J\left(g^J - g^{J-1}\right), & 2^J \leq T \\ 0, & \text{otherwise} \end{cases}, \quad (3)$$

The next lemma details its properties (cf. Dorfman & Levy, 2022, Lemma 3.2; and Appendix B).

Lemma 3.1. *For $\mathcal{M}_f$ satisfying Equation (2), we have that*

1. $\text{Bias}(g^{\text{MLMC}}) := \left\|\mathbb{E}g^{\text{MLMC}} - \nabla f(x)\right\| \leq \sqrt{2c^2/T}$.
2. $\text{Var}(g^{\text{MLMC}}) := \mathbb{E}\left\|g^{\text{MLMC}} - \mathbb{E}g^{\text{MLMC}}\right\|^2 \leq 14c^2\log T$.
3. *The expected cost of constructing g^{MLMC} is $\mathcal{O}(\log T)$.*

This result implies that we can use $\mathcal{M}_f$ to construct a gradient estimator with: **(1)** low bias, proportional to $1/\sqrt{T}$; **(2)** near-constant variance; and **(3)** only logarithmic cost.

3.3. Byzantine-Robustness with MLMC Gradients

Next, we show that combining the MLMC estimator with a robust aggregation rule ensures Byzantine-robustness. We consider (δ, κ_δ)-robust aggregation rules, a concept recently introduced by Allouah et al. (2023), which unifies preceding formulations like $(\delta_{\max}, c)$-*RAgg* (Karimireddy et al., 2022).

Definition 3.2 ((δ, κ_δ)-robustness)**.** Let $\delta < 1/2$ and $\kappa_\delta \geq 0$. An aggregation rule $\mathcal{A}$ is (δ, κ_δ)-robust if for any vectors $g_1, ..., g_m$, and any set $S \subseteq [m]$ of size $(1-\delta)m$, we have

$$\left\|\mathcal{A}(g_1, \dots, g_m) - \overline{g}_S\right\|^2 \leq \frac{\kappa_\delta}{|S|}\sum_{i\in S}\left\|g_i - \overline{g}_S\right\|^2,$$

where $\overline{g}_S = \frac{1}{|S|}\sum_{i\in S} g_i$.

This definition includes, for example, coordinate-wise median (CWMed, Yin et al., 2018) and geometric median (Pillutla et al., 2022). The lemma below shows that robustly aggregating mini-batch gradients from honest workers yields a mapping satisfying Equation (2) with $c^2 = 2\sigma^2\left(\kappa_\delta + \frac{1}{m}\right)$.

Lemma 3.3. *Consider $x \in \mathcal{K}$ and let $\overline{g}_1^N, \dots, \overline{g}_m^N$ be m vectors such that $\forall i \in \mathcal{G}$, $\overline{g}_i^N$ is a mini-batch gradient estimator based on N i.i.d samples, i.e., $\overline{g}_i^N = \frac{1}{N}\sum_{n=1}^N \nabla F(x; \xi_i^n)$, where $\xi_i^n \overset{i.i.d}{\sim} \mathcal{D}$. Then, under Assumption 2.1, any (δ, κ_δ)-robust aggregation rule $\mathcal{A}$ satisfies*

$$\mathbb{E}\left\|\mathcal{A}(\overline{g}_1^N, \dots, \overline{g}_m^N) - \nabla f(x)\right\|^2 \leq \frac{2\sigma^2}{N}\left(\kappa_\delta + \frac{1}{m}\right).$$

Algorithm 1 Byzantine-Robust Optimization with MLMC

Input: Initial iterate $x_1 \in \mathbb{R}^d$, learning rate η.
for $t = 1, \ldots, T$ **do**
  Draw $J_t \sim \text{Geom}(1/2)$
  for $i \in [m]$ in parallel **do**
    for $j \in \{0, J_t - 1, J_t\}$ **do**
      Compute $\bar{g}_{t,i}^j \leftarrow \frac{1}{2^j} \sum_{k=1}^{2^j} \nabla F(x_t; \xi_{t,i}^k)$
    end for
    Send $(\bar{g}_{t,i}^0, \bar{g}_{t,i}^{J_t-1}, \bar{g}_{t,i}^{J_t})$ if $i \in \mathcal{G}$, else send $(*, *, *)$
  end for
  for $j \in \{0, J_t - 1, J_t\}$ **do**
    $\widehat{g}_t^j \leftarrow \mathcal{A}(\bar{g}_{t,1}^j, \ldots, \bar{g}_{t,m}^j)$ // robust aggregation
  end for
  $g_t \leftarrow \widehat{g}_t^0 + \begin{cases} 2^{J_t}(\widehat{g}_t^{J_t} - \widehat{g}_t^{J_t-1}), & J_t \leq J_{\max} = \lfloor \log T \rfloor \\ 0, & \text{otherwise.} \end{cases}$
  $x_{t+1} \leftarrow \Pi_{\mathcal{K}}(x_t - \eta g_t)$
end for

Building on this result, we propose Algorithm 1. In each round $t \in [T]$, honest workers compute and send mini-batch gradients of sizes 1, 2^{J_t-1}, and 2^{J_t}, with $J_t \sim \text{Geom}(1/2)$. Then, the server applies a robust aggregation rule to each group of gradients and subsequently constructs an MLMC estimator as in Equation (3) to perform an SGD update. The next result confirms the convergence of Algorithm 1 for non-convex functions; its full proof and convex analysis are deferred to Appendix C, with a proof sketch provided here.

Theorem 3.4. *Under Assumption 2.1, with a (δ, κ_δ)-robust aggregator $\mathcal{A}$, consider Algorithm 1 with learning rate given by $\eta = \min\left\{\frac{\sqrt{\Delta_1}}{4\sigma\sqrt{L\gamma T \log T}}, \frac{1}{L}\right\}$, where $\gamma := \kappa_\delta + \frac{1}{m}$ and $\Delta_1 := f(x_1) - f^*$. Denoting $\nabla_t := \nabla f(x_t)$, it holds that*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla_t\|^2 \leq 16\sqrt{\frac{L\Delta_1\sigma^2\gamma\log T}{T}} + \frac{2\big(L\Delta_1 + 2\sigma^2\gamma\big)}{T}.$$

When $\kappa_\delta \in \mathcal{O}(\delta)$, e.g., for CWMed combined with Nearest Neighbor Mixing (Allouah et al., 2023), the rate in Theorem 3.4 is consistent with the state-of-the-art convergence guarantees (Karimireddy et al., 2021; Allouah et al., 2023), up to a $\sqrt{\log T}$ factor. Moreover, the expected per-worker sample complexity is $\mathcal{O}(T \log T)$, representing a modest increase of only a $\log T$ factor over existing methods. Yet, our approach is conceptually different that prior work. Instead of implicitly reducing Byzantine-induced bias through honest worker variance reduction, we use direct bias reduction strategy by constructing MLMC gradients post-aggregation.

Proof Sketch. For $\eta \leq \frac{1}{L}$, one can show that the SGD updates $x_{t+1} = x_t - \eta g_t$ satisfy (cf. Lemma A.2)

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla_t\|^2 \leq \frac{2\Delta_1}{T\eta} + \frac{\eta L}{T}\sum_{t=1}^{T}\mathbb{E}V_t^2 + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|b_t\|^2 , \tag{4}$$

where $b_t := \mathbb{E}g_t - \nabla_t$ and $V_t^2 := \mathbb{E}\|g_t - \mathbb{E}g_t\|^2$ are the bias and variance of g_t, respectively. Combining Lemma 3.1 with Lemma 3.3 implies that for every $t \in [T]$, we have

$$\|b_t\| \leq 2\sigma\sqrt{\frac{\gamma}{T}}, \quad \text{and} \quad V_t^2 \leq 28\sigma^2\gamma\log T . \tag{5}$$

Plugging these bounds and tuning η concludes the proof. □

4. DynaBRO: Dynamic Byzantine-Robustness

In this section, we transition to the dynamic setting, demonstrating how a slightly modified version of Algorithm 1 endures a non-trivial number of rounds with ID relabeling (i.e., identity changes). This modification involves adding a fail-safe filter designed to mitigate potential bias from identity switches during the MLMC gradient construction.

Since the MLMC gradient in round t depends solely on 2^{J_t} computations per-worker in that round, if identity changes occur only between rounds, e.g., when only communications could be altered, then Algorithm 1 can be seamlessly applied in the dynamic setting and our analysis remains valid. However, Byzantine-robustness addresses a broader range of failures, where identity changes may also arise during different gradient computations within a round. For instance, in data poisoning attacks (Huang et al., 2020; Schwarzschild et al., 2021), some gradients for the same worker might be contaminated while others remain clean, depending on the integrity of the sampled data. To address this challenge, we propose a fail-safe filter specifically designed for the MLMC gradient. Initially, we slightly adjust our notation: we denote by $\mathcal{G}_t^k$ the set of honest workers in round t, at the k-th gradient computation, with $k \in \left[2^{J_t}\right]$. In addition, we define $\tau_s := \left\{t : \mathcal{G}_t^1 = \cdots = \mathcal{G}_t^{2^{J_t}}\right\}$ as the set of *static* rounds, where worker identities remain fixed within the round, and $\tau_d := [T] \setminus \tau_s$ indicates the set of *dynamic* rounds.

MLMC fail-safe filter. Recall the MLMC gradient formula for $J_t \leq \lfloor \log T \rfloor$, $g_t = \widehat{g}_t^0 + 2^{J_t}(\widehat{g}_t^{J_t} - \widehat{g}_t^{J_t-1})$. Its bias-variance analysis in the static case hinges on Lemma 3.3, which asserts that for each level j, the squared distance between $\widehat{g}_t^j$ and $\widehat{g}_t^{j-1}$ is proportional to 2^{-j}. However, this lemma is not applicable in dynamic rounds due to the absence of a consistent set of honest workers. Since the MLMC elements of a worker, $\bar{g}_{t,i}^{J_t-1}$ and $\bar{g}_{t,i}^{J_t}$, are averages of gradients computed during the round, even a single instance of Byzantine behavior could compromise them. This might disrupt the expected trend of decreasing distances between aggregated gradients at consecutive levels. To counteract

Algorithm 2 DynaBRO (**Dyna**mic **B**yzantine-**R**obust **O**ptimization)

Input: Initial iterate $x_1 \in \mathbb{R}^d$, learning rate sequence $\{\eta_t\}_{t\geq 1}$, universal coefficient $C := \sqrt{8\log(16m^2T)}$.
for $t = 1, \ldots, T$ **do**
  Draw $J_t \sim \text{Geom}(1/2)$
  for $k = 1, \ldots, 2^{J_t}$ **do**
    for $i \in [m]$ in parallel **do**
      Compute and send $g_{t,i}^k \leftarrow \nabla F(x_t; \xi_{t,i}^k)$ if $i \in \mathcal{G}_t^k$, else send $*$
    end for
  end for
  for $i \in [m]$ in parallel **do**
    Compute $\bar{g}_{t,i}^j \leftarrow \frac{1}{2^j}\sum_{k=1}^{2^j} g_{t,i}^k$ for $j \in \{0, J_t - 1, J_t\}$
  end for
  for $j \in \{0, J_t - 1, J_t\}$ **do**
    Option 1: $\widehat{g}_t^j \leftarrow \mathcal{A}(\bar{g}_{t,1}^j, \ldots, \bar{g}_{t,m}^j)$, $c_\mathcal{E} := \sqrt{\gamma}$ // $\mathcal{A}$ is (δ, κ_δ)-robust; $\gamma := 2\kappa_\delta + \frac{1}{m}$
    Option 2: $\widehat{g}_t^j \leftarrow \text{MFM}(\bar{g}_{t,1}^j, \ldots, \bar{g}_{t,m}^j; \mathcal{T}^j := 2C\mathcal{V}/\sqrt{2^j})$, $c_\mathcal{E} := 6\sqrt{2}$ // See Algorithm 3
  end for
  Define fail-safe event $\mathcal{E}_t = \left\{\|\widehat{g}_t^{J_t} - \widehat{g}_t^{J_t-1}\| \leq (1+\sqrt{2})c_\mathcal{E} C\mathcal{V}/\sqrt{2^{J_t}}\right\}$
  Construct MLMC gradient $g_t \leftarrow \widehat{g}_t^0 + \begin{cases} 2^{J_t}(\widehat{g}_t^{J_t} - \widehat{g}_t^{J_t-1}), & \text{if } J_t \leq \lfloor \log T \rfloor \text{ and } \mathcal{E}_t \text{ holds} \\ 0, & \text{otherwise.} \end{cases}$
  Update $x_{t+1} \leftarrow \Pi_\mathcal{K}(x_t - \eta_t g_t)$
end for

this potential manipulation, we introduce an event $\mathcal{E}_t$ to verify that this distance remains within expected bounds:

$$\mathcal{E}_t = \left\{\left\|\widehat{g}_t^{J_t} - \widehat{g}_t^{J_t-1}\right\| \leq (1+\sqrt{2})c_\mathcal{E} C\mathcal{V}/\sqrt{2^{J_t}}\right\} , \quad (6)$$

where $C := \sqrt{8\log(16m^2T)}$. If $\mathcal{E}_t$ holds, we proceed with the standard MLMC construction; otherwise, we revert to a simpler aggregated gradient using a single sample per-worker, similar to when $J_t > \lfloor \log T \rfloor$ (see Algorithm 2). The parameter $c_\mathcal{E}$ is set to ensure that $\mathcal{E}_t$ holds with high probability in static rounds, where this modification contributes only a lower-order term to the bias. In dynamic rounds, it restricts the bias to a near-constant level.

The convergence of our modified algorithm is established in the following theorem, detailed in Appendix D.

Theorem 4.1. *Under Assumption 2.2, with a (δ, κ_δ)-robust aggregator $\mathcal{A}$, consider Algorithm 2 with **Option 1** and with a fixed learning rate $\eta := \min\left\{\frac{\sqrt{\Delta_1}}{3C\mathcal{V}\sqrt{L\gamma T\log T}}, \frac{1}{L}\right\}$, where $\gamma := 2\kappa_\delta + \frac{1}{m}$ and $\Delta_1 := f(x_1) - f^*$. Then, it holds that*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(x_t)\|^2 \leq 12C\mathcal{V}\sqrt{\frac{L\Delta_1\gamma\log T}{T}} + \frac{2L\Delta_1 + 9C^2\mathcal{V}^2\gamma}{T} + 16C^2\mathcal{V}^2\gamma\frac{|\tau_d|\log T}{T} ,$$

where $C := \sqrt{8\log(16m^2T)}$.

When there are no identity switches during gradient computations within the same round (i.e., $|\tau_d| = 0$), we revert to the rate established in Theorem 3.4 for the static setting, but with σ effectively replaced by $C\mathcal{V}$ due to differing noise assumptions. The bounded noise assumption allows the application of a concentration inequality, ensuring that $\mathcal{E}_t$ occurs with high probability in static rounds. When identity switches are present, convergence is still assured, provided that the number of identity-switching rounds is sub-linear. Specifically, Theorem 4.1 leads to the subsequent corollary.

Corollary 4.2. *The asymptotic convergence rate implied by Theorem 4.1 is given by*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(x_t)\|^2 \in \tilde{\mathcal{O}}\left(\mathcal{V}\sqrt{\frac{L\Delta_1\gamma}{T}} + \mathcal{V}^2\gamma\frac{|\tau_d|}{T}\right) .$$

Thus, Algorithm 2 can withstand $|\tau_d| \in \tilde{\mathcal{O}}(\sqrt{T/\gamma})$ dynamic rounds (omitting dependence on L, Δ_1, and $\mathcal{V}$), while matching the static convergence rate. Concretely, when $\kappa_\delta \in \mathcal{O}(\delta)$, this rate becomes $\tilde{\mathcal{O}}\left(\mathcal{V}\sqrt{L\Delta_1(\delta + 1/m)/T}\right)$.

Proof Sketch. Following the methodology of Theorem 3.4, we bound $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(x_t)\|^2$ as in Equation (4). The bias and variance bounds for g_t are detailed in Lemma D.4. We highlight key differences from the static setting: in static rounds, the bias is similarly bounded with an additional lower-order term, reflecting instances where $\mathcal{E}_t$ does not hold. In dynamic rounds, $\mathcal{E}_t$ limits the expected distance be-

tween aggregated gradients at consecutive levels as follows:

$$\mathbb{E}_{t-1}\left[\|\widehat{g}_t^j - \widehat{g}_t^{j-1}\|^2 \cdot \mathbb{1}_{\mathcal{E}_t(j)}\right] = \mathbb{E}_{t-1}\left[\|\widehat{g}_t^j - \widehat{g}_t^{j-1}\|^2 | \mathcal{E}_t(j)\right] \mathbb{P}(\mathcal{E}_t(j)) \in \tilde{\mathcal{O}}\left(\frac{\mathcal{V}^2\gamma}{2^j}\right) ,$$

where $\mathcal{E}_t(j)$ is defined as in Equation (6) given $J_t = j$. Taking expectation w.r.t J_t leads to

$$\text{MSE}_t \coloneqq \mathbb{E}_{t-1}\|g_t - \nabla f(x_t)\|^2 \in \tilde{\mathcal{O}}(\mathcal{V}^2\gamma) .$$

Consequently, the bias and variance are bounded as $\tilde{\mathcal{O}}(\mathcal{V}\sqrt{\gamma})$ and $\tilde{\mathcal{O}}(\mathcal{V}^2\gamma)$, respectively. Substituting these bounds and setting η completes the proof. □

When worker-momentum fails. As detailed in Appendix E, the worker-momentum approach may fail when Byzantine workers change identities. By meticulously crafting a momentum-tailored *dynamic attack* that leverages the momentum parameter and exploits its diminishing effect on past gradients, we can induce sufficient drift (i.e., bias) across all momentums simultaneously. This strategy requires only $\mathcal{O}(\sqrt{T})$ rounds of identity switches, which our method can withstand, as shown in Corollary 4.2. It remains an intriguing open question whether the worker-momentum method can be augmented with a mechanism similar to our fail-safe filter to handle identity changes.

5. Optimality and Adaptivity

While we have demonstrated convergence for Algorithm 2 when employing a general (δ, κ_δ)-robust aggregator, this class of aggregators does not ensure optimal convergence under the bounded noise assumption. This limitation arises because the most effective aggregator features $\kappa_\delta \in \mathcal{O}(\delta)$ (Allouah et al., 2023), suggesting that the Byzantine-related error term scales with $\sqrt{\delta/T}$ at best. Conversely, the optimal scaling under this noise assumption is proportional to $\delta/\sqrt{T}$ (Alistarh et al., 2018; Allen-Zhu et al., 2020).

In this section, we introduce the *Median-Filtered Mean* (MFM) aggregator. Although it does not meet the (δ, κ_δ)-robustness criteria, it facilitates near-optimal convergence rates. Additionally, by employing an adaptive learning rate, we eliminate the need for prior knowledge of the smoothness parameter and the fraction of Byzantine workers, which is typically necessary to determine the learning rate.

Median-Filtered Mean. Consider vectors $g_1, \ldots, g_m$, and a threshold parameter $\mathcal{T}$. Our proposed aggregation method, outlined in Algorithm 3, computes the mean of vectors within $\mathcal{T}$-proximity of a representative median vector, g_{med}. This median is chosen to ensure that the majority of messages fall within a $\mathcal{T}/2$ radius of it. If no such median vector exists, i.e., if there is no vector around which at least half of the other messages lie within a $\mathcal{T}/2$ radius, the algorithm defaults to outputting the zero vector. We note that a similar mechanism for gradient estimation was previously used by Alistarh et al. (2018); Allen-Zhu et al. (2020).

Algorithm 3 Median-Filtered Mean (MFM)

Input: Vectors $g_1, \ldots, g_m$, threshold $\mathcal{T}$.
$\mathcal{M} \leftarrow \{i \in [m] : |\{j \in [m] : \|g_j - g_i\| \leq \mathcal{T}/2\}| > m/2\}$
if $\mathcal{M} \neq \emptyset$ **then**
  $g_{\text{med}} \leftarrow g_i$ for some $i \in \mathcal{M}$
  $\widehat{\mathcal{G}} \leftarrow \{i \in [m] : \|g_i - g_{\text{med}}\| \leq \mathcal{T}\}$
  $\widehat{g} \leftarrow \frac{1}{|\widehat{\mathcal{G}}|}\sum_{i \in \widehat{\mathcal{G}}} g_i$
else
  $\widehat{g} \leftarrow 0$
end if
Return: $\widehat{g}$

In Appendix F.1, we demonstrate that the MFM aggregator does not satisfy the (δ, κ_δ)-robustness criteria. Yet, by appropriately setting the threshold parameter $\mathcal{T}$, we achieve a superior asymptotic bound on the distance between aggregated and true gradients in static rounds. This leads to an improved convergence rate, as we detail later. What follows is an informal statement of Lemma F.3.

Lemma 5.1 (Informal)**.** *Consider the setting of Lemma 3.3, replacing Assumption 2.1 with Assumption 2.2, and let $\widehat{g}^N = \mathit{MFM}(\bar{g}_1^N, \ldots, \bar{g}_m^N; \mathcal{T}^N)$ be the output of Algorithm 3 with $\mathcal{T}^N \in \tilde{\Theta}(\mathcal{V}/\sqrt{N})$. Then, with high probability,*

$$\left\|\widehat{g}^N - \nabla f(x)\right\|^2 \in \tilde{\mathcal{O}}\left(\frac{\mathcal{V}^2}{N}\left(\delta^2 + \frac{1}{m}\right)\right) .$$

Conversely, a (δ, κ_δ)-robust aggregator yields a worse asymptotic bound of $\tilde{\mathcal{O}}\left(\frac{\mathcal{V}^2}{N}(\kappa_\delta + \frac{1}{m})\right)$; see Lemma D.1.

Adaptive learning rate. For some $\eta_0 > 0$, we consider the following version of the AdaGrad-Norm learning rate (Levy et al., 2018; Ward et al., 2020; Faw et al., 2022):

$$\eta_t = \frac{\eta_0}{\sqrt{\sum_{s=1}^t \|g_s\|^2}} . \tag{7}$$

It is well-known that AdaGrad-Norm adapts to the gradient's variance and the objective's smoothness, achieving the same asymptotic convergence rates as if these parameters were known in advance (Kavis et al., 2022; Attia & Koren, 2023). Unlike the specifically tuned learning rate in Theorem 4.1, this adaptive learning rate does not incorporate $\mathcal{V}$, L, or δ. Yet, our method still requires knowledge of the noise level $\mathcal{V}$ to set the MFM threshold. While it may initially seem that δ is also necessary for configuring the parameter $c_\mathcal{E}$ within the event $\mathcal{E}_t$, we can adjust $c_\mathcal{E}$ to be independent of δ by trivially

bounding it. Thus, our method effectively adapts to both the smoothness L and the fraction of Byzantine workers δ.

We now present the convergence result for Algorithm 2 when employing the MFM aggregator (**Option 2**) alongside the AdaGrad-Norm learning rate. For ease of analysis, we consider problems with bounded objectives, such as neural networks with bounded output activations, e.g., sigmoid or softmax (cf. Appendix H for a detailed analysis).

Theorem 5.2. *Suppose Assumption 2.2 holds and f is bounded by M (i.e., $\max_x |f(x)| \leq M$). Considering Algorithm 2 with **Option 2** and the AdaGrad-Norm learning rate as given in Equation (7), define $\zeta := \frac{2M}{\eta_0} + \eta_0 L$. Then,*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left\|\nabla f(x_t)\right\|^2 \in \tilde{\mathcal{O}}\left(\zeta\mathcal{V}\sqrt{\frac{\tilde{\gamma}}{T}} + \frac{\zeta^2}{T} + \mathcal{V}^2\frac{|\tau_d|}{T}\right) ,$$

where $\tilde{\gamma} := 32\delta^2 + \frac{1}{m}$.

In contrast to Theorem 4.1, the third term, associated with the number of dynamic rounds, lacks a factor of $\tilde{\gamma}$ due to our adjustment of the parameter $c_{\mathcal{E}}$ to $\mathcal{O}(1)$ instead of $\mathcal{O}(\sqrt{\tilde{\gamma}})$. Had we utilized the latter, as in **Option 1**, a similar bound could be achieved, but it would necessitate prior knowledge of δ. Consequently, the absence of this $\tilde{\gamma}$ factor restricts the number of dynamic rounds we can withstand without compromising convergence to $|\tau_d| \in \tilde{\mathcal{O}}(\sqrt{\tilde{\gamma}T})$, a more restrictive bound compared to Corollary 4.2, as detailed in Corollary H.5. With this more restrictive number of dynamic rounds, we achieve a (near-)optimal convergence rate of $\tilde{\mathcal{O}}\left(\mathcal{V}\sqrt{(\delta^2 + 1/m)/T}\right)$. This observation raises a compelling open question: Does adaptivity inherently lead to decreased robustness against Byzantine identity changes?

Proof Sketch. Our proof mirrors the approach used in Theorem 4.1, with two key differences. First, using the AdaGrad-Norm learning rate leads to $G_T^2 := \sum_{t=1}^{T} \|\nabla f(x_t)\|^2$ exhibiting a 'self-boundness' property, in contrast to the non-adaptive, learning rate-dependent bound in Equation (4). Lemma G.4 formalizes this property, indicating:

$$\mathbb{E}G_T^2 \leq 2\zeta\left(\sqrt{\sum_{t=1}^{T}\mathbb{E}V_t^2} + \sqrt{2\mathbb{E}S_T^2} + \sqrt{2\mathbb{E}G_T^2}\right) + \mathbb{E}S_T^2 ,$$

where $S_T^2 := \sum_{t=1}^{T} \|b_t\|^2$. Secondly, adjusting $c_{\mathcal{E}}$ to $\mathcal{O}(1)$ results in slightly different bias and variance bounds in dynamic rounds, lacking $\tilde{\gamma}$ factors. Specifically, we have:[4]

$$\|b_t\| \in \begin{cases} \tilde{\mathcal{O}}(\mathcal{V}), & t \in \tau_d \\ \tilde{\mathcal{O}}(\mathcal{V}\sqrt{\frac{\tilde{\gamma}}{T}}), & t \in \tau_s \end{cases}, \quad V_t^2 \in \begin{cases} \tilde{\mathcal{O}}(\mathcal{V}^2), & t \in \tau_d \\ \tilde{\mathcal{O}}(\mathcal{V}^2\gamma), & t \in \tau_s \end{cases}.$$

Applying these bounds, solving for $\mathbb{E}G_T^2$, and dividing by T yields the final bound. □

[4]Here, we ignore the low-probability event where the MFM aggregator outputs zero. See Lemma H.2 for a formal statement.

6. Experiments

In this section, we provide numerical experiments to evaluate our approach. In our experiments, we aim to demonstrate: **(1)** the trade-off between an algorithm's history window size dependence and its susceptibility to dynamic Byzantine changes; and **(2)** the benefit of our MLMC-based method compared to the prominent worker-momentum approach (Karimireddy et al., 2021; Farhadkhani et al., 2022; Allouah et al., 2023). To this end, we study two types of simulated dynamic identity-switching strategies:

1. **Periodic(K):** Once in every K rounds, a new subset (of size δm) of Byzantine workers is sampled uniformly at random. Between any two such samplings, all worker identities remain fixed. A lower value of K corresponds to a higher rate of identity switches, which implies a stronger dynamic attack.

2. **Bernoulli($p, D, \delta_{\max}$):** For each worker, we sample $X \sim \text{Ber}(p)$ independently across workers and iterations. If $X = 1$, then the worker becomes Byzantine for a fixed duration of D iterations, up to a maximum of $\delta_{\max}$-fraction of Byzantine workers per iteration.

Note that for the **Periodic** strategy, the number of Byzantine workers in each iteration remains the same, whereas for the **Bernoulli** strategy, it changes throughout training.

We study image classification on the MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky et al., 2009) datasets using convolutional neural networks (CNNs) with 2 and 4 layers, respectively, as in Allouah et al. (2023). Additional training details are deferred to Appendix J for brevity. We benchmark our method against worker-momentum using momentum parameter $\beta \in \{0.9, 0.99\}$, as well as against vanilla SGD, which corresponds to $\beta = 0$. In our experiments, we did not use the fail-safe MLMC filter, opting instead for Algorithm 1. Since the MLMC estimator typically requires multiple gradient computations per update, to ensure a fair comparison, we present all results based on an equivalent total number of gradient computations. We ran all experiments with 5 random seeds and report their mean and standard deviation.

MNIST. We first consider the MNIST dataset under the **Periodic**(K) switching strategy for different values of $K \in \{5, 10, 20, 100\}$ and $K \to \infty$, corresponding to the static setting where the set of Byzantine workers is fixed. In this setup, we consider $m = 17$ workers, of which $\delta m = 8$ are Byzantine. In Figure 1, we visualize the final test accuracy when the Byzantine workers implement the sign-flip attack (SF, Allen-Zhu et al., 2020) and the server uses the coordinate-wise trimmed mean aggregator (CWTM, Yin et al., 2018). As observed, the performance of our approach

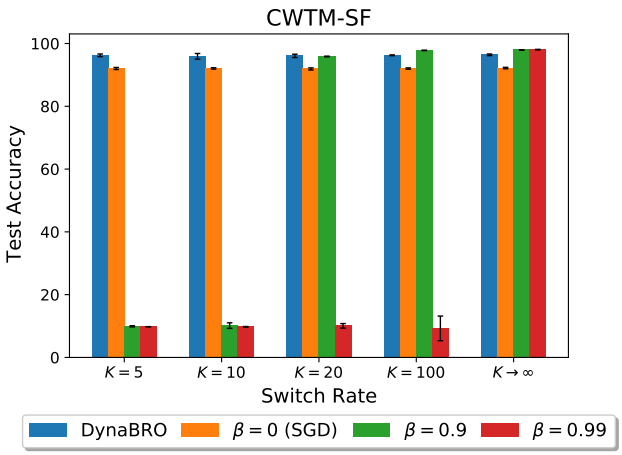

Figure 1. Final test accuracy on MNIST under the **Periodic**(K) identity-switching strategy for different values of K. Byzantine workers implement the SF attack and the server employs CWTM.

remains stable across different values of K. On the other hand, for worker-momentum, performance declines as K decreases, with higher momentum values experiencing more significant effects. For example, momentum with $\beta = 0.99$ fails when $K = 100$, whereas with $\beta = 0.9$ it performs well even when $K = 20$. These results are expected, as momentum with parameter β effectively averages the last $\frac{1}{1-\beta}$ iterations, leading us to anticipate this switching rate as its 'break-off' point. In Appendix J.1, we provide the results of an analogous experiment with a different pair of attack and aggregator, demonstrating a similar trend.

CIFAR-10. Next, we consider CIFAR-10 classification with $m = 25$ workers under the **Bernoulli**($p, D, \delta_{\max}$) switching strategy. We investigate three switching configurations: **(1)** $p = 0.01, D = 10$; **(2)** $p = 0.01, D = 50$; and **(3)** $p = 0.05, D = 10$. For all configurations, we restrict the maximum fraction of Byzantine workers in any single iteration to $\delta_{\max} = 0.72$ (equivalent to 18 out of 25 workers), indicating that the fraction of Byzantine workers in any given iteration may exceed 0.5. In Appendix J.2, we present similar results for $\delta_{\max} = 0.48$. The Byzantine workers employ the inner-product manipulation attack (IPM, Xie et al., 2020), while the server utilizes CWMed. We compare our method against vanilla SGD and momentum with parameter $\beta = 0.9$. In Figure 2, we display the test accuracy and histograms of the fraction of Byzantine workers throughout training. Surprisingly, for the first configuration, which exhibits a relatively low number of Byzantine workers in any iteration, momentum with $\beta = 0.9$ slightly outperforms our method. However, in the other configurations, which have a larger number of Byzantine workers per iteration and a non-negligible number of iterations with $\delta > 0.5$, our method significantly outperforms both SGD and momentum.

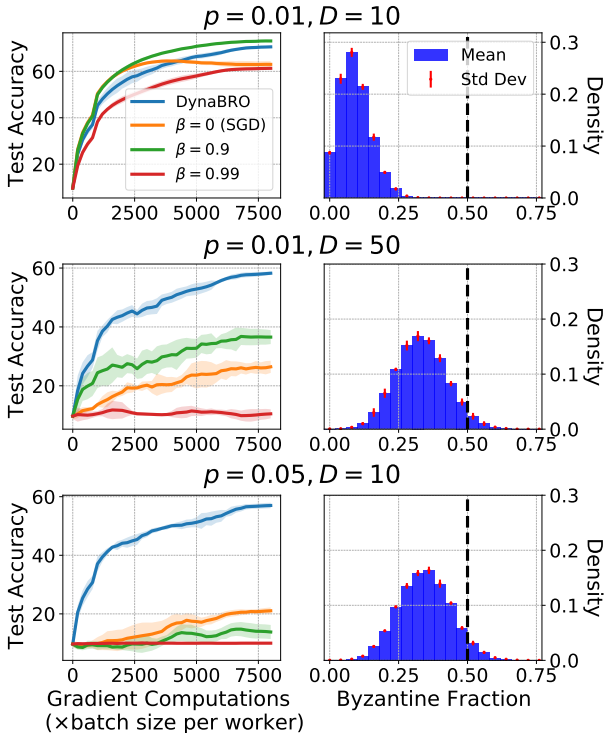

Figure 2. Test accuracy and histogram of the fraction of Byzantine workers over time on CIFAR-10 under the **Bernoulli**($p, D, \delta_{\max}$) identity-switching strategy for different values of p and D. Byzantine workers employ the IPM attack and the server uses CWMed.

7. Conclusion and Future Work

In this work, we introduced DynaBRO, a novel approach for Byzantine-robust learning in dynamic settings. We tackled the challenge of Byzantine behavior alterations, demonstrating that our method withstands a substantial number of Byzantine identity changes while achieving the asymptotic convergence rate of the static setting. A key innovation is the use of an MLMC gradient estimation technique and its integration with a fail-safe filter, which enhances robustness against dynamic Byzantine strategies. Coupled with an adaptive learning rate, our approach further alleviates the necessity for prior knowledge of the smoothness parameter and the fraction of Byzantine workers.

Several important avenues for future research emerge from our study. First, we study the homogeneous case, where all workers minimize the same objective. Extending our analysis to heterogeneous datasets is an exciting and challenging direction since a direct application of our approach results in suboptimal bounds for this setting. Another direction includes exploring simultaneous adaptivity to the noise level, the smoothness, and the fraction of Byzantine workers; this presents a complex challenge as we are unaware of any optimal aggregation rule that is agnostic to both the noise level and the fraction of Byzantine workers.

Acknowledgments

This research was partially supported by Israel PBC-VATAT, by the Technion Artificial Intelligent Hub (Tech.AI) and by the Israel Science Foundation (grant No. 447/20).

Impact Statement

This paper introduces advancements in Byzantine-robust learning. While inherent societal implications concerning security and reliability in distributed ML are acknowledged, we do not feel they must be specifically highlighted here.

References

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

Alistarh, D., Allen-Zhu, Z., and Li, J. Byzantine stochastic gradient descent. *Advances in Neural Information Processing Systems*, 31, 2018.

Allen-Zhu, Z., Ebrahimian, F., Li, J., and Alistarh, D. Byzantine-resilient non-convex stochastic gradient descent. *arXiv preprint arXiv:2012.14368*, 2020.

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

Asi, H., Carmon, Y., Jambulapati, A., Jin, Y., and Sidford, A. Stochastic bias-reduced gradient methods. *Advances in Neural Information Processing Systems*, 34:10810–10822, 2021.

Atre, M., Jha, B., and Rao, A. Distributed deep learning using volunteer computing-like paradigm. In *2021 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pp. 933–942. IEEE, 2021.

Attia, A. and Koren, T. SGD with AdaGrad Stepsizes: Full Adaptivity with High Probability to Unknown Parameters, Unbounded Gradients and Affine Variance. *arXiv preprint arXiv:2302.08783*, 2023.

Auer, P., Cesa-Bianchi, N., and Gentile, C. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002.

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

Beznosikov, A., Samsonov, S., Sheshukova, M., Gasnikov, A., Naumov, A., and Moulines, E. First Order Methods with Markovian Noise: from Acceleration to Variational Inequalities. *arXiv preprint arXiv:2305.15938*, 2023.

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

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečnỳ, J., Mazzocchi, S., McMahan, B., et al. Towards federated learning at scale: System design. *Proceedings of machine learning and systems*, 1:374–388, 2019.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

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

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

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

Dorfman, R. and Levy, K. Y. Adapting to mixing time in stochastic optimization with markovian data. In *International Conference on Machine Learning*, pp. 5429–5446. PMLR, 2022.

El-Mhamdi, E.-M., Guerraoui, R., and Rouault, S. Distributed momentum for byzantine-resilient learning. *arXiv preprint arXiv:2003.00010*, 2020.

Farhadkhani, S., Guerraoui, R., Gupta, N., Pinot, R., and Stephan, J. Byzantine machine learning made easy by resilient averaging of momentums. In *International Conference on Machine Learning*, pp. 6246–6283. PMLR, 2022.

Faw, M., Tziotis, I., Caramanis, C., Mokhtari, A., Shakkottai, S., and Ward, R. The power of adaptivity in sgd: Self-tuning step sizes with unbounded gradients and affine variance. In *Conference on Learning Theory*, pp. 313–355. PMLR, 2022.

Feng, J., Xu, H., and Mannor, S. Distributed robust learning. *arXiv preprint arXiv:1409.5937*, 2014.

Giles, M. B. Multilevel monte carlo methods. *Acta numerica*, 24:259–328, 2015.

Gorbunov, E., Horváth, S., Richtárik, P., and Gidel, G. Variance Reduction is an Antidote to Byzantines: Better Rates, Weaker Assumptions and Communication Compression as a Cherry on the Top. *arXiv preprint arXiv:2206.00529*, 2022.

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

Guerraoui, R., Gupta, N., and Pinot, R. Byzantine machine learning: A primer. *ACM Computing Surveys*, 2023.

Hu, Y., Chen, X., and He, N. On the bias-variance-cost tradeoff of stochastic optimization. *Advances in Neural Information Processing Systems*, 34:22119–22131, 2021.

Huang, W. R., Geiping, J., Fowl, L., Taylor, G., and Goldstein, T. Metapoison: Practical general-purpose clean-label data poisoning. *Advances in Neural Information Processing Systems*, 33:12080–12091, 2020.

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

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

Karimireddy, S. P., He, L., and Jaggi, M. Byzantine-robust learning on heterogeneous datasets via bucketing. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=jXKKDEi5vJt.

Kavis, A., Levy, K., and Cevher, V. High probability bounds for a class of nonconvex algorithms with AdaGrad stepsize. In *10th International Conference on Learning Representations (ICLR)*, 2022.

Kijsipongse, E., Piyatumrong, A., and U-ruekolan, S. A hybrid GPU cluster and volunteer computing platform for scalable deep learning. *The Journal of Supercomputing*, 74:3236–3263, 2018.

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

Lamport, L., Shostak, R., and Pease, M. The Byzantine generals problem. pp. 203–226. 2019.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33: 8847–8860, 2020.

Levy, K. Online to offline conversions, universality and adaptive minibatch sizes. *Advances in Neural Information Processing Systems*, 30, 2017.

Levy, K. Y., Yurtsever, A., and Cevher, V. Online adaptive methods, universality and acceleration. *Advances in neural information processing systems*, 31, 2018.

Malinovsky, G., Richtárik, P., Horváth, S., and Gorbunov, E. Byzantine Robustness and Partial Participation Can Be Achieved Simultaneously: Just Clip Gradient Differences. *arXiv preprint arXiv:2311.14127*, 2023.

McMahan, H. B. and Streeter, M. Adaptive bound optimization for online convex optimization. *arXiv preprint arXiv:1002.4908*, 2010.

Mohri, M. and Yang, S. Accelerating online convex optimization via adaptive prediction. In *Artificial Intelligence and Statistics*, pp. 848–856. PMLR, 2016.

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

Pinelis, I. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, pp. 1679–1706, 1994.

Ryabinin, M. and Gusev, A. Towards crowdsourced training of large neural networks using decentralized mixture-of-experts. *Advances in Neural Information Processing Systems*, 33:3659–3672, 2020.

Schwarzschild, A., Goldblum, M., Gupta, A., Dickerson, J. P., and Goldstein, T. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In *International Conference on Machine Learning*, pp. 9389–9398. PMLR, 2021.

Shi, Y. and Cornish, R. On multilevel Monte Carlo unbiased gradient estimation for deep latent variable models. In *International Conference on Artificial Intelligence and Statistics*, pp. 3925–3933. PMLR, 2021.

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

Suttle, W. A., Bedi, A., Patel, B., Sadler, B. M., Koppel, A., and Manocha, D. Beyond exponentially fast mixing in average-reward reinforcement learning via multi-level Monte Carlo actor-critic. In *International Conference on Machine Learning*, pp. 33240–33267. PMLR, 2023.

Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T., and Rellermeyer, J. S. A survey on distributed machine learning. *Acm computing surveys (csur)*, 53(2): 1–33, 2020.

Ward, R., Wu, X., and Bottou, L. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *The Journal of Machine Learning Research*, 21(1):9047–9076, 2020.

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

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

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

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

Appendix Outline

We provide a concise outline of the appendix structure as follows:

Appendix A. General analysis of SGD with biased gradients for convex and non-convex functions.

Appendix B. Proof of properties for the MLMC estimator (Lemma 3.1).

Appendix C. Analysis of the MLMC approach in the static Byzantine setting for convex and non-convex objectives.

Appendix D. Analysis of the dynamic Byzantine setting where identities change over time.

Appendix E. Analyzing how the worker-momentum method fails in scenarios involving Byzantine ID relabeling.

Appendix F. Properties of the MFM aggregator, introduced to guarantee optimal convergence rates.

Appendix G. Results mirroring Appendix A using AdaGrad-Norm as the learning rate.

Appendix H. Convergence guarantees for DynaBRO with MFM and AdaGrad-Norm – improved rates and adaptivity.

Appendix I. Technical lemmata.

Appendix J. Experimental setup and training details.

A. SGD with Biased Gradients

Consider the SGD update rule, defined for some initial iterate $x_1 \in \mathbb{R}^d$ and a fixed learning rate $\eta > 0$ as,

$$x_{t+1} = \Pi_{\mathcal{K}}(x_t - \eta g_t) \ , \tag{8}$$

where g_t is an estimator of $\nabla f(x_t)$ with bias $b_t \coloneqq \mathbb{E}[g_t - \nabla f(x_t)|x_t]$ and variance $V_t^2 \coloneqq \mathbb{E}[\|g_t - \mathbb{E}g_t\|^2 \, |x_t]$.

The following lemmas establish bounds on the optimality gap and sum of squared gradients norm for SGD with (possibly) biased gradients, when applied to convex and non-convex functions, respectively. All our convergence results rely on these lemmas; to be precise, we derive bounds on the bias and variance of the relevant gradient estimator and plug them into our results in a black-box fashion.

Lemma A.1 (Convex SGD)**.** *Assume f is convex. Consider Equation (8) with $\eta \leq \frac{1}{2L}$. If the domain $\mathcal{K}$ is bounded with diameter D (Assumption 2.3), then*

$$\mathbb{E}f(\bar{x}_T) - f(x^*) \leq \frac{D^2}{2T\eta} + \frac{\eta}{T}\sum_{t=1}^{T}\mathbb{E}V_t^2 + \frac{D}{T}\sum_{t=1}^{T}\mathbb{E}\left\|b_t\right\| \ ,$$

where $\bar{x}_T \coloneqq \frac{1}{T}\sum_{t=2}^{T+1} x_t$.

Proof. By the convexity of f, the gradient inequality implies that

$$\sum_{t\in[T]}(\mathbb{E}f(x_t) - f(x^*)) \leq \sum_{t\in[T]}\mathbb{E}[\nabla f(x_t)^\top(x_t - x^*)] = \sum_{t\in[T]}\mathbb{E}[g_t^\top(x_t - x^*)] - \sum_{t\in[T]}\mathbb{E}[b_t^\top(x_t - x^*)] \ . \tag{9}$$

Focusing on the first term in the R.H.S, by applying Lemma I.2 with $x = x^*$, we have

$$\begin{aligned}\sum_{t\in[T]}\mathbb{E}[g_t^\top(x_t - x^*)] &\leq \frac{1}{2\eta}\sum_{t\in[T]}\left(\mathbb{E}\|x_t - x^*\|^2 - \mathbb{E}\|x_{t+1} - x^*\|^2\right) + \sum_{t\in[T]}\left(\mathbb{E}[g_t^\top(x_t - x_{t+1})] - \frac{1}{2\eta}\mathbb{E}\|x_t - x_{t+1}\|^2\right) \\ &\leq \frac{D^2}{2\eta} + \sum_{t\in[T]}\left(\mathbb{E}[g_t^\top(x_t - x_{t+1})] - \frac{1}{2\eta}\mathbb{E}\|x_t - x_{t+1}\|^2\right) \ ,\end{aligned} \tag{10}$$

where the last inequality follows from a telescoping sum and $\|x_1 - x^*\|^2 \leq D^2$. On the other hand, by the smoothness of f, we can bound the L.H.S as follows:

$$\sum_{t\in[T]} (\mathbb{E}f(x_t) - f(x^*)) \geq \sum_{t\in[T]} \left(\mathbb{E}f(x_{t+1}) - f(x^*) - \mathbb{E}[\nabla f(x_t)^\top (x_{t+1} - x_t)] - \frac{L}{2}\mathbb{E}\|x_t - x_{t+1}\|^2 \right) . \tag{11}$$

Plugging Equations (10) and (11) back into Equation (9) yields:

$$\begin{aligned}
&\sum_{t\in[T]} (\mathbb{E}f(x_{t+1}) - f(x^*)) \leq \\
&\qquad \frac{D^2}{2\eta} + \sum_{t\in[T]} \left(\mathbb{E}[(g_t - \nabla f(x_t))^\top (x_t - x_{t+1})] - \left(\frac{1}{2\eta} - \frac{L}{2}\right) \mathbb{E}\|x_t - x_{t+1}\|^2 \right) - \sum_{t\in[T]} \mathbb{E}[b_t^\top (x_t - x^*)] \\
&\leq \frac{D^2}{2\eta} + \sum_{t\in[T]} \left(\mathbb{E}[(g_t - \nabla f(x_t))^\top (x_t - x_{t+1})] - \frac{1}{4\eta}\mathbb{E}\|x_t - x_{t+1}\|^2 \right) - \sum_{t\in[T]} \mathbb{E}[b_t^\top (x_t - x^*)] \\
&\leq \frac{D^2}{2\eta} + \sum_{t\in[T]} \left(\mathbb{E}[(g_t - \mathbb{E}g_t)^\top (x_t - x_{t+1})] - \frac{1}{4\eta}\mathbb{E}\|x_t - x_{t+1}\|^2 \right) - \sum_{t\in[T]} \mathbb{E}[b_t^\top (x_{t+1} - x^*)] \\
&\leq \frac{D^2}{2\eta} + \eta \sum_{t\in[T]} \mathbb{E}V_t^2 - \sum_{t\in[T]} \mathbb{E}[b_t^\top (x_{t+1} - x^*)] ,
\end{aligned}$$

where the second inequality uses $\frac{1}{2\eta} - \frac{L}{2} \geq \frac{1}{4\eta}$ and the final inequality uses Young's inequality, $a^\top b - \frac{1}{2}\|b\|^2 \leq \frac{1}{2}\|a\|^2$. Using Cauchy-Schwarz inequality, we have $-b_t^\top (x_{t+1} - x^*) \leq \|b_t\| \|x_{t+1} - x^*\| \leq D\|b_t\|$; plugging this bound and using Jensen's inequality gives

$$\mathbb{E}f(\bar{x}_T) - f(x^*) \leq \frac{1}{T}\sum_{t\in[T]} \mathbb{E}f(x_{t+1}) - f(x^*) \leq \frac{D^2}{2T\eta} + \frac{\eta}{T}\sum_{t\in[T]} \mathbb{E}V_t^2 + \frac{D}{T}\sum_{t\in[T]} \mathbb{E}\|b_t\| .$$

□

Lemma A.2 (Non-convex SGD)**.** *Consider Equation (8) with $\mathcal{K} = \mathbb{R}^d$ (i.e., unconstrained) and $\eta \leq \frac{1}{L}$, and let $\Delta_1 := f(x_1) - f^*$. Then,*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla f(x_t)\|^2 \leq \frac{2\Delta_1}{T\eta} + \frac{\eta L}{T}\sum_{t=1}^{T} \mathbb{E}V_t^2 + \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|b_t\|^2 .$$

Proof. We begin our proof by following the methodology presented in Lemma 2 of (Ajalloeian & Stich, 2020). By the smoothness of f, we have

$$\begin{aligned}
\mathbb{E}_{t-1}f(x_{t+1}) &\leq f(x_t) - \eta\nabla f(x_t)^\top \mathbb{E}_{t-1}g_t + \frac{\eta^2 L}{2}\mathbb{E}_{t-1}\|g_t\|^2 \\
&= f(x_t) - \eta\nabla f(x_t)^\top (\nabla f(x_t) + b_t) + \frac{\eta^2 L}{2}\left(V_t^2 + \mathbb{E}_{t-1}\|\mathbb{E}g_t\|^2\right) \\
&= f(x_t) - \eta\|\nabla f(x_t)\|^2 - \eta\nabla f(x_t)^\top b_t + \frac{\eta^2 L}{2}V_t^2 + \frac{\eta^2 L}{2}\|\nabla f(x_t) + b_t\|^2 \\
&\leq f(x_t) - \eta\|\nabla f(x_t)\|^2 - \eta\nabla f(x_t)^\top b_t + \frac{\eta^2 L}{2}V_t^2 + \frac{\eta}{2}\left(\|\nabla f(x_t)\|^2 + 2\nabla f(x_t)^\top b_t + \|b_t\|^2\right) \\
&= f(x_t) - \frac{\eta}{2}\left(\|\nabla f(x_t)\|^2 + \|b_t\|^2\right) + \frac{\eta^2 L}{2}V_t^2 ,
\end{aligned}$$

where the last inequality follows from $\eta \leq \frac{1}{L}$. Denote: $\Delta_t := \mathbb{E}f(x_t) - f^*$. By rearranging terms and taking expectation, we obtain

$$\mathbb{E}\|\nabla f(x_t)\|^2 \leq \frac{2(\Delta_t - \Delta_{t+1})}{\eta} + \eta L\mathbb{E}V_t^2 + \mathbb{E}\|b_t\|^2 .$$

Summing over $t \in [T]$,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\|\nabla f(x_t)\right\|^2 \leq \frac{2}{T\eta}\sum_{t=1}^{T}(\Delta_t - \Delta_{t+1}) + \frac{\eta L}{T}\sum_{t=1}^{T}\mathbb{E}V_t^2 + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\|b_t\right\|^2$$
$$\leq \frac{2\Delta_1}{T\eta} + \frac{\eta L}{T}\sum_{t=1}^{T}\mathbb{E}V_t^2 + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\|b_t\right\|^2 ,$$

which concludes the proof. □

B. Efficient Bias-Reduction with MLMC Gradient Estimation

In this section, we establish the properties of the MLMC estimator in Equation (3) through the proof of Lemma 3.1.

Lemma 3.1. *For $\mathcal{M}_f$ satisfying Equation (2), we have that*

1. $\text{Bias}(g^{\text{MLMC}}) := \left\|\mathbb{E}g^{\text{MLMC}} - \nabla f(x)\right\| \leq \sqrt{2c^2/T}$.
2. $\text{Var}(g^{\text{MLMC}}) := \mathbb{E}\left\|g^{\text{MLMC}} - \mathbb{E}g^{\text{MLMC}}\right\|^2 \leq 14c^2\log T$.
3. *The expected cost of constructing g^{MLMC} is $\mathcal{O}(\log T)$.*

Proof. Our proof follows those presented in Lemma 3.1 of (Dorfman & Levy, 2022) and in Proposition 1 of (Asi et al., 2021). Let $J_{\max} = \lfloor \log T \rfloor$, and recall that $\mathbb{E}\left\|g^j - \nabla f(x)\right\|^2 \leq \frac{c^2}{2^j}$. By explicitly writing the expectation over $J \sim \text{Geom}(\frac{1}{2})$, we have

$$\mathbb{E}g^{\text{MLMC}} = \mathbb{E}g^0 + \sum_{j=1}^{J_{\max}} 2^{-j}\cdot 2^j\mathbb{E}[g^j - g^{j-1}] = \mathbb{E}[g^{J_{\max}}] ,$$

where the last equality follows from a telescoping sum. Thus, be Jensen's inequality, it holds that

$$\left\|\mathbb{E}g^{\text{MLMC}} - \nabla f(x)\right\| = \left\|\mathbb{E}g^{J_{\max}} - \nabla f(x)\right\| \leq \sqrt{\mathbb{E}\|g^{J_{\max}} - \nabla f(x)\|^2} \leq \sqrt{\frac{c^2}{2^{J_{\max}}}} \leq \sqrt{\frac{2c^2}{T}} ,$$

where the last inequality follows from $2^{J_{\max}} = 2^{\lfloor \log T \rfloor} \geq T/2$. For the second part, we have

$$\mathbb{E}\left\|g^{\text{MLMC}} - \mathbb{E}g^{\text{MLMC}}\right\|^2 \leq \mathbb{E}\left\|g^{\text{MLMC}} - \nabla f(x)\right\|^2 \leq 2\mathbb{E}\left\|g^{\text{MLMC}} - g^0\right\|^2 + 2\mathbb{E}\left\|g^0 - \nabla f(x)\right\|^2 .$$

Focusing on the first term in the R.H.S,

$$\mathbb{E}\left\|g^{\text{MLMC}} - g^0\right\|^2 = \sum_{j=1}^{J_{\max}} 2^{-j}\cdot 2^{2j}\mathbb{E}\left\|g^j - g^{j-1}\right\|^2$$
$$\leq 2\sum_{j=1}^{J_{\max}} 2^j\left(\mathbb{E}\left\|g^j - \nabla f(x)\right\|^2 + \mathbb{E}\left\|g^{j-1} - \nabla f(x)\right\|^2\right)$$
$$\leq 2\sum_{j=1}^{J_{\max}} 2^j\left(\frac{c^2}{2^j} + \frac{c^2}{2^{j-1}}\right) = 6\sum_{j=1}^{J_{\max}} c^2 \leq 6c^2\log T ,$$

where the last inequality uses $J_{\max} \leq \log T$. Using $\mathbb{E}\left\|g^0 - \nabla f(x)\right\|^2 \leq c^2$, we have

$$\mathbb{E}\left\|g^{\text{MLMC}} - \mathbb{E}g^{\text{MLMC}}\right\|^2 \leq 2\cdot 6c^2\log T + 2\cdot c^2 \leq 14c^2\log T .$$

Finally, since we call $\mathcal{M}_f(x;1)$, and with probability 2^{-j} we call $\mathcal{M}_f(x;2^j)$ and $\mathcal{M}_f(x;2^{j-1})$, the expected number of stochastic gradient evaluations is at most $1 + \sum_{j=1}^{J_{\max}} 2^{-j}\left(2^j + 2^{j-1}\right) = 1 + \frac{3}{2}J_{\max} \leq 1 + \frac{3}{2}\log T \in \mathcal{O}(\log T)$. □

C. Static Byzantine-Robustness with MLMC Gradient Estimation

In this section, we analyze Algorithm 1 in the standard, static setting, where the identity of Byzantine workers remains fixed. We show that distributed SGD with MLMC estimation applied to robustly-aggregated gradients is Byzantine-resilient.

We begin by establishing Lemma 3.3, which asserts that utilizing a (δ, κ_δ)-robust aggregator when honest workers compute stochastic gradients over a mini-batch satisfies Equation (2) with $c^2 = 2\sigma^2\left(\kappa_\delta + \frac{1}{m}\right)$.

Lemma 3.3. *Consider $x \in \mathcal{K}$ and let $\bar{g}_1^N, \ldots, \bar{g}_m^N$ be m vectors such that $\forall i \in \mathcal{G}$, $\bar{g}_i^N$ is a mini-batch gradient estimator based on N i.i.d samples, i.e., $\bar{g}_i^N = \frac{1}{N}\sum_{n=1}^N \nabla F(x;\xi_i^n)$, where $\xi_i^n \overset{i.i.d}{\sim} \mathcal{D}$. Then, under Assumption 2.1, any (δ, κ_δ)-robust aggregation rule $\mathcal{A}$ satisfies*

$$\mathbb{E}\left\|\mathcal{A}(\bar{g}_1^N, \ldots, \bar{g}_m^N) - \nabla f(x)\right\|^2 \leq \frac{2\sigma^2}{N}\left(\kappa_\delta + \frac{1}{m}\right) .$$

Proof. For ease of notation, denote $\widehat{g}^N \coloneqq \mathcal{A}(\bar{g}_1^N, \ldots, \bar{g}_m^N)$ and $\bar{g}^N \coloneqq \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\bar{g}_i^N$. Since $\mathbb{E}\bar{g}^N = \nabla f(x)$, and by the (δ, κ_δ)-robustness of $\mathcal{A}$, we have

$$\begin{aligned}
\mathbb{E}\left\|\widehat{g}^N - \nabla f(x)\right\|^2 &= \mathbb{E}\left\|\widehat{g}^N - \bar{g}^N\right\|^2 + \mathbb{E}\left\|\bar{g}^N - \nabla f(x)\right\|^2 \\
&\leq \frac{\kappa_\delta}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left\|\bar{g}_i^N - \bar{g}^N\right\|^2 + \mathbb{E}\left\|\bar{g}^N - \nabla f(x)\right\|^2 \\
&= \frac{\kappa_\delta}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mathbb{E}\left\|\bar{g}_i^N - \nabla f(x)\right\|^2 + (\kappa_\delta + 1)\mathbb{E}\left\|\bar{g}^N - \nabla f(x)\right\|^2 \\
&\leq \kappa_\delta\frac{\sigma^2}{N} + (\kappa_\delta + 1)\frac{\sigma^2}{|\mathcal{G}|N} \\
&= \frac{\sigma^2}{N}\left(\kappa_\delta + \frac{\kappa_\delta + 1}{|\mathcal{G}|}\right) \\
&\leq \frac{2\sigma^2}{N}\left(\kappa_\delta + \frac{1}{m}\right) ,
\end{aligned}$$

where the second inequality holds as $\bar{g}_i^N$ for every $i \in \mathcal{G}$ and $\bar{g}^N$ are the averages of N and $|\mathcal{G}|N$ i.i.d samples, respectively, each with variance bounded by σ^2. In the last inequality we used $|\mathcal{G}| > m/2$. □

C.1. Convex Case

Next, we now establish the convergence of Algorithm 1 in the convex setting.

Theorem C.1. *Assume f is convex. Under Assumptions 2.1 and 2.3, and with a (δ, κ_δ)-robust aggregator $\mathcal{A}$, consider Algorithm 1 with learning rate*

$$\eta = \min\left\{\frac{D}{8\sigma\sqrt{\gamma T \log T}}, \frac{1}{2L}\right\} ,$$

where $\gamma \coloneqq \kappa_\delta + \frac{1}{m}$. Then, for every $T \geq 1$,

$$\mathbb{E}f(\bar{x}_T) - f(x^*) \leq 10D\sigma\sqrt{\frac{\gamma\log T}{T}} + \frac{LD^2}{T} .$$

Proof. Employing Lemma A.1, we have

$$\mathbb{E}f(\bar{x}_T) - f(x^*) \leq \frac{D^2}{2T\eta} + \frac{\eta}{T}\sum_{t=1}^T \mathbb{E}V_t^2 + \frac{D}{T}\sum_{t=1}^T \mathbb{E}\left\|b_t\right\| , \tag{12}$$

where b_t and V_t are the bias and variance of g_t, respectively. Combining Lemma 3.1 with Lemma 3.3 implies that

$$\|b_t\| = \|\mathbb{E}g_t - \nabla f(x_t)\| \leq \sqrt{\frac{2\cdot 2\sigma^2\gamma}{T}} = 2\sigma\sqrt{\frac{\gamma}{T}}, \quad \text{and} \quad \mathbb{E}V_t^2 = \mathbb{E}\|g_t - \mathbb{E}g_t\|^2 \leq 28\sigma^2\gamma\log T . \tag{13}$$

We note that computing g_t requires $\mathcal{O}(m \log T)$ stochastic gradient evaluations, in expectation. Plugging these bounds back to Equation (12) gives:

$$\mathbb{E}f(\bar{x}_T) - f(x^*) \leq \frac{D^2}{2T\eta} + 28\eta\sigma^2\gamma\log T + 2D\sigma\sqrt{\frac{\gamma}{T}} \leq \frac{1}{2}\left(\frac{D^2}{T\eta} + 64\eta\sigma^2\gamma\log T\right) + 2D\sigma\sqrt{\frac{\gamma}{T}} \ .$$

Since $\eta = \min\left\{\frac{D}{8\sigma\sqrt{\gamma T\log T}}, \frac{1}{2L}\right\}$, applying Lemma I.7 with $a = \frac{D^2}{T}, b = 64\sigma^2\gamma\log T$, and $c = 2L$, enables to bound the sum of the first two terms as,

$$\frac{D^2}{T\eta} + 64\eta\sigma^2\gamma\log T \leq 16D\sigma\sqrt{\frac{\gamma\log T}{T}} + \frac{2LD^2}{T} \ .$$

Plugging this bound back gives:

$$\mathbb{E}f(\bar{x}_T) - f(x^*) \leq 8D\sigma\sqrt{\frac{\gamma\log T}{T}} + \frac{LD^2}{T} + 2D\sigma\sqrt{\frac{\gamma}{T}} \leq 10D\sigma\sqrt{\frac{\gamma\log T}{T}} + \frac{LD^2}{T} \ .$$

□

C.2. Non-convex Case

Moving forward, we establish the convergence of Algorithm 1 for non-convex functions in Theorem 3.4, restated here.

Theorem 3.4. *Under Assumption 2.1, with a (δ, κ_δ)-robust aggregator $\mathcal{A}$, consider Algorithm 1 with learning rate given by $\eta = \min\left\{\frac{\sqrt{\Delta_1}}{4\sigma\sqrt{L\gamma T\log T}}, \frac{1}{L}\right\}$, where $\gamma := \kappa_\delta + \frac{1}{m}$ and $\Delta_1 := f(x_1) - f^*$. Denoting $\nabla_t := \nabla f(x_t)$, it holds that*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla_t\|^2 \leq 16\sqrt{\frac{L\Delta_1\sigma^2\gamma\log T}{T}} + \frac{2\left(L\Delta_1 + 2\sigma^2\gamma\right)}{T}.$$

Proof. We follow the proof of Theorem C.1, substituting Lemma A.1 with Lemma A.2, which implies that

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\|\nabla f(x_t)\right\|^2 \leq \frac{2\Delta_1}{T\eta} + \frac{\eta L}{T}\sum_{t=1}^{T}\mathbb{E}V_t^2 + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\|b_t\right\|^2 \ . \tag{14}$$

Plugging the bounds in Equation (13), we then obtain

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\|\nabla f(x_t)\right\|^2 \leq \frac{2\Delta_1}{T\eta} + 28\eta L\sigma^2\gamma\log T + \frac{4\sigma^2\gamma}{T} \leq 2\left(\frac{\Delta_1}{T\eta} + 16\eta L\sigma^2\gamma\log T\right) + \frac{4\sigma^2\gamma}{T} \ .$$

Since $\eta = \min\left\{\frac{\sqrt{\Delta_1}}{4\sigma\sqrt{L\gamma T\log T}}, \frac{1}{L}\right\}$, applying Lemma I.7 with $a = \frac{\Delta_1}{T}, b = 16L\sigma^2\gamma\log T$, and $c = L$, allows us to bound the sum of the first two terms as follows

$$\frac{\Delta_1}{T\eta} + 16\eta L\sigma^2\gamma\log T \leq 8\sqrt{\frac{L\Delta_1\sigma^2\gamma\log T}{T}} + \frac{L\Delta_1}{T} \ .$$

Substituting this bound back yields:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\|\nabla f(x_t)\right\|^2 \leq 16\sqrt{\frac{L\Delta_1\sigma^2\gamma\log T}{T}} + \frac{2\left(L\Delta_1 + 2\sigma^2\gamma\right)}{T} \ .$$

□

D. Dynamic Byzantine-Robustness with General (δ, κ)-robust Aggregator

In this section, we analyze DynaBRO (Algorithm 2) with **Option 1**, which utilizes a general (δ, κ)-robust aggregator.

Recall that Algorithm 2 with **Option 1** performs the following update rule for every $t \in [T]$:

$$J_t \sim \text{Geom}(1/2)$$

$$\widehat{g}_t^j \leftarrow \mathcal{A}(\bar{g}_{t,1}^j, \ldots, \bar{g}_{t,m}^j), \quad \text{where } \bar{g}_{t,i}^j = \frac{1}{2^j}\sum_{k=1}^{2^j} \nabla F(x_t; \xi_{t,i}^k) \text{ for every } i \in \mathcal{G}_t \text{ if } t \notin \tau_d \tag{15}$$

$$g_t \leftarrow \widehat{g}_t^0 + \begin{cases} 2_t^J \left(\widehat{g}_t^{J_t} - \widehat{g}_t^{J_t-1}\right), & \text{if } J_t \leq J_{\max} := \lfloor \log T \rfloor \text{ and } \mathcal{E}_t(J_t) \text{ holds} \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

$$x_{t+1} \leftarrow \Pi_{\mathcal{K}}(x_t - \eta_t g_t) \ ,$$

where the associated event $\mathcal{E}_t(J_t)$ in this scenario is defined as,

$$\mathcal{E}_t(J_t) := \left\{ \|\widehat{g}_t^{J_t} - \widehat{g}_t^{J_t-1}\| \leq (1+\sqrt{2})\frac{c_{\mathcal{E}} C \mathcal{V}}{\sqrt{2^{J_t}}} \right\}, \quad c_{\mathcal{E}} := \sqrt{\gamma}, \quad C := \sqrt{8\log(16m^2T)}, \quad \gamma := 2\kappa_\delta + \frac{1}{m} \ . \tag{17}$$

We start by establishing a deterministic and a high probability bound on the distance between the true gradient and the robustly-aggregated stochastic gradients, when honest workers compute gradients over a mini-batch. By combining these bounds and adjusting the probability parameter, we provide an upper bound on the expected squared distance, i.e., MSE, as presented in Corollary D.2.

Lemma D.1. *Consider the setting in Lemma 3.3, i.e., let $x \in \mathcal{K}$ and $\bar{g}_1^N, \ldots, \bar{g}_m^N$ be m vectors such that for each $i \in \mathcal{G}$, $\bar{g}_i^N$ is a mini-batch gradient estimator based on $N \in \mathbb{N}$ i.i.d samples. Then, under Assumption 2.2, any (δ, κ_δ)-robust aggregation rule $\mathcal{A}$ satisfies,*

1. $\left\|\mathcal{A}(\bar{g}_1^N, \ldots, \bar{g}_m^N) - \nabla f(x)\right\|^2 \leq 2\left(4\kappa_\delta + 1\right)\mathcal{V}^2$.
2. *With probability at least $1-p$,*
$$\left\|\mathcal{A}(\bar{g}_1^N, \ldots, \bar{g}_m^N) - \nabla f(x)\right\|^2 \leq C_p^2 \frac{\mathcal{V}^2\gamma}{N}$$
where $C_p^2 = 8\log(4m/p)$.

Proof. Denote the aggregated gradient and the empirical average of honest workers by $\widehat{g}^N := \mathcal{A}(\bar{g}_1^N, \ldots, \bar{g}_m^N)$ and $\bar{g}^N := \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} \bar{g}_i^N$, respectively. Thus,

$$\begin{aligned}
\left\|\widehat{g}^N - \nabla f(x)\right\|^2 &\leq 2\left\|\widehat{g}^N - \bar{g}^N\right\|^2 + 2\left\|\bar{g}^N - \nabla f(x)\right\|^2 \\
&\leq \frac{2\kappa_\delta}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\bar{g}_i^N - \bar{g}^N\right\|^2 + 2\left\|\bar{g}^N - \nabla f(x)\right\|^2 \\
&\leq \frac{4\kappa_\delta}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left\|\bar{g}_i^N - \nabla\right\|^2 + (4\kappa_\delta + 2)\left\|\bar{g}^N - \nabla f(x)\right\|^2
\end{aligned}$$

where we used $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ and the (δ, κ_δ)-robustness of $\mathcal{A}$. Since it trivially holds that $\left\|\bar{g}_i^N - \nabla f(x)\right\| \leq \mathcal{V}$ for every $i \in \mathcal{G}$ and $\left\|\bar{g}^N - \nabla f(x)\right\| \leq \mathcal{V}$, we have

$$\left\|\widehat{g}^N - \nabla f(x)\right\|^2 \leq 2(4\kappa_\delta + 1)\mathcal{V}^2 \ ,$$

which establishes the first part. For the second part, we employ the concentration argument presented in Lemma I.1. With probability at least $1-\tilde{p}$, it holds that

$$\left\|\bar{g}^N - \nabla f(x)\right\| \leq \mathcal{V}\sqrt{\frac{2\log(2/\tilde{p})}{N|\mathcal{G}|}} \leq 2\mathcal{V}\sqrt{\frac{\log(2/\tilde{p})}{Nm}} \ , \tag{18}$$

as $|\mathcal{G}| > m/2$. Additionally, for each $i \in \mathcal{G}$ separately, we have with probability at least $1 - \tilde{p}$ that

$$\left\|\bar{g}_i^N - \nabla f(x)\right\| \leq \mathcal{V}\sqrt{\frac{2\log(2/\tilde{p})}{N}} \ . \tag{19}$$

Hence, with probability at least $1 - (1 + |\mathcal{G}|)\tilde{p}$, the union bound ensures that Equations (18) and (19) hold simultaneously for every $i \in \mathcal{G}$. Since $1 - (1 + |\mathcal{G}|)\tilde{p} \geq 1 - 2|\mathcal{G}|\tilde{p} \geq 1 - 2m\tilde{p}$, with probability at least $1 - p$ it holds that

$$\begin{aligned}\left\|\widehat{g}^N - \nabla f(x)\right\|^2 \leq 4\kappa_\delta \left(\mathcal{V}\sqrt{\frac{2\log(4m/p)}{N}}\right)^2 + (4\kappa_\delta + 2)\left(2\mathcal{V}\sqrt{\frac{\log(4m/p)}{Nm}}\right)^2 &= 8\log\left(\frac{4m}{p}\right)\frac{\mathcal{V}^2}{N}\left(\kappa_\delta + \frac{2\kappa_\delta + 1}{m}\right) \\ &\leq 8\log\left(\frac{4m}{p}\right)\frac{\mathcal{V}^2}{N}\left(2\kappa_\delta + \frac{1}{m}\right) \ ,\end{aligned}$$

which concludes the proof, assuming $m \geq 2$. □

Corollary D.2 (MSE of Aggregated Gradients)**.** *Let $\bar{g}_1^N, \ldots, \bar{g}_m^N$ be as defined in Lemma D.1, with $N \leq T$. Under Assumption 2.2, any (δ, κ_δ)-robust aggregator $\mathcal{A}$ satisfies,*

$$\mathbb{E}\left\|\mathcal{A}(\bar{g}_1^N, \ldots, \bar{g}_m^N) - \nabla f(x)\right\|^2 \leq \frac{2C^2\mathcal{V}^2\gamma}{N} \ ,$$

where $C = \sqrt{8\log(16m^2T)}$ and $\gamma = 2\kappa_\delta + \frac{1}{m}$ as in Equation (17).

Proof. By choosing $p = \frac{1}{4mT}$, item 2 of Lemma D.1 implies that with probability at least $1 - \frac{1}{4mT}$,

$$\left\|\mathcal{A}(\bar{g}_1^N, \ldots, \bar{g}_m^N) - \nabla f(x)\right\|^2 \leq \frac{C^2\mathcal{V}^2\gamma}{N} \ .$$

In addition, by item 1 of Lemma D.1, we have $\left\|\mathcal{A}(\bar{g}_1^N, \ldots, \bar{g}_m^N) - \nabla f(x)\right\|^2 \leq 2(4\kappa_\delta + 1)\mathcal{V}^2$, deterministically. Combining these results, by the law of total expectation, we get

$$\mathbb{E}\left\|\mathcal{A}(\bar{g}_1^N, \ldots, \bar{g}_m^N) - \nabla f(x)\right\|^2 \leq \frac{C^2\mathcal{V}^2\gamma}{N} + 2(4\kappa_\delta + 1)\mathcal{V}^2 \cdot \frac{1}{4mT} \leq \frac{C^2\mathcal{V}^2\gamma}{N} + \frac{\mathcal{V}^2\gamma}{T} \leq \frac{2C^2\mathcal{V}^2\gamma}{N} \ ,$$

where the second inequality follows from $\frac{4\kappa_\delta + 1}{2m} \leq \gamma$, and the last inequality from $C^2 \geq 1$ and $N \leq T$. □

Before we establish bounds on the bias and variance of the MLMC gradient estimator defined in Equation (16), we show that $\mathcal{E}_t$ is satisfied with high probability.

Lemma D.3. *Consider $\mathcal{E}(J_t)$ defined in Equation (17). For every $t \in \tau_s$ and $j = 0, \ldots, J_{\max}$, we have*

$$\mathbb{P}_{t-1}(\mathcal{E}_t(j)) \geq 1 - \frac{1}{2mT} \ ,$$

where the randomness is w.r.t the stochastic gradient samples.

Proof. By item 2 of Lemma D.1, we have

$$\mathbb{P}_{t-1}\left(\|\widehat{g}_t^j - \nabla_t\| \leq \frac{c_\mathcal{E} C\mathcal{V}}{\sqrt{2^j}}\right) = \mathbb{P}_{t-1}\left(\|\widehat{g}_t^j - \nabla_t\| \leq C\mathcal{V}\sqrt{\frac{\gamma}{2^j}}\right) \geq 1 - \frac{1}{4mT}, \quad \forall j = 0, \ldots, J_{\max} \ .$$

This bound, in conjunction with the union bound, allows us to bound $\mathbb{P}_{t-1}(\mathcal{E}_t(j)^c)$ as,

$$\begin{aligned}\mathbb{P}_{t-1}(\mathcal{E}_t(j)^c) &= \mathbb{P}_{t-1}\left(\|\widehat{g}_t^j - \widehat{g}_t^{j-1}\| > (1 + \sqrt{2})\frac{c_\mathcal{E} C\mathcal{V}}{\sqrt{2^j}}\right) \\ &\leq \mathbb{P}_{t-1}\left(\left\{\|\widehat{g}_t^j - \nabla_t\| > \frac{c_\mathcal{E} C\mathcal{V}}{\sqrt{2^j}}\right\} \bigcup \left\{\|\widehat{g}_t^{j-1} - \nabla_t\| > \frac{c_\mathcal{E} C\mathcal{V}}{\sqrt{2^{j-1}}}\right\}\right) \\ &\leq \mathbb{P}_{t-1}\left(\|\widehat{g}_t^j - \nabla_t\| > \frac{c_\mathcal{E} C\mathcal{V}}{\sqrt{2^j}}\right) + \mathbb{P}_{t-1}\left(\|\widehat{g}_t^{j-1} - \nabla_t\| > \frac{c_\mathcal{E} C\mathcal{V}}{\sqrt{2^{j-1}}}\right) \\ &\leq \frac{1}{4mT} + \frac{1}{4mT} = \frac{1}{2mT} \ .\end{aligned}$$

□

Moving forward, we now provide bounds on the bias and variance estimator in Equation (16). In Lemma D.4, we establish that the bias in static rounds ($t \in \tau_s$) is proportionate to $\mathcal{V}\sqrt{\gamma/T}$, whereas in bad rounds it is near-constant; the variance is also near-constant for every $t \in [T]$.

Lemma D.4 (MLMC Bias and Variance)**.** *Consider g_t defined as in Equation (16). Then,*

1. *The bias $b_t \coloneqq \mathbb{E}_{t-1} g_t - \nabla_t$ is bounded as*

$$\|b_t\| \leq \begin{cases} 2C\mathcal{V}\sqrt{\frac{\gamma}{T}} + 2\mathcal{V}\sqrt{\frac{\gamma}{m}}\frac{\log T}{T}, & t \in \tau_s \\ 4C\mathcal{V}\sqrt{\gamma \log T}, & t \in \tau_d \end{cases} .$$

2. *The variance $V_t^2 \coloneqq \mathbb{E}_{t-1} \|g_t - \mathbb{E}_{t-1} g_t\|^2$ is bounded as*

$$V_t^2 \leq 16C^2\mathcal{V}^2\gamma \log T, \quad \forall t \in [T] .$$

Proof. Our proof closely follows and builds upon the strategy employed in Lemma 3.1. We begin by bounding the variance,

$$\begin{aligned} V_t^2 \coloneqq \mathbb{E}_{t-1} \|g_t - \mathbb{E}_{t-1} g_t\|^2 \leq \mathbb{E}_{t-1} \|g_t - \nabla_t\|^2 &= \sum_{j=1}^{\infty} 2^{-j} \mathbb{E}_{t-1} \left\| \widehat{g}_t^0 + 2^j \left(\widehat{g}_t^j - \widehat{g}_t^{j-1} \right) \mathbb{1}_{\{j \leq J_{\max}\} \cap \mathcal{E}_t(j)} - \nabla_t \right\|^2 \\ &\leq 2 \sum_{j=1}^{\infty} 2^{-j} \mathbb{E}_{t-1} \left\| \widehat{g}_t^0 - \nabla_t \right\|^2 + 2 \sum_{j=1}^{J_{\max}} 2^j \mathbb{E}_{t-1} \left[\|\widehat{g}_t^j - \widehat{g}_t^{j-1}\|^2 \mathbb{1}_{\mathcal{E}_t(j)} \right] \\ &= 2\mathbb{E}_{t-1} \left\| \widehat{g}_t^0 - \nabla_t \right\|^2 + 2 \sum_{j=1}^{J_{\max}} 2^j \underbrace{\mathbb{E}_{t-1} \left[\|\widehat{g}_t^j - \widehat{g}_t^{j-1}\|^2 \mathbb{1}_{\mathcal{E}_t(j)} \right]}_{=(\dagger)} , \end{aligned} \tag{20}$$

where the last equality holds as $\sum_{j=1}^{\infty} 2^{-j} = 1$. Focusing on $(\dagger)$, by the law of total expectation, we have that

$$\mathbb{E}_{t-1} \left[\|\widehat{g}_t^j - \widehat{g}_t^{j-1}\|^2 \mathbb{1}_{\mathcal{E}_t(j)} \right] = \mathbb{E}_{t-1} \left[\|\widehat{g}_t^j - \widehat{g}_t^{j-1}\|^2 | \mathcal{E}_t(j) \right] \underbrace{\mathbb{P}(\mathcal{E}_t(j))}_{\leq 1} \leq \frac{(1+\sqrt{2})^2 C^2 \mathcal{V}^2 \gamma}{2^j} \leq \frac{6C^2\mathcal{V}^2\gamma}{2^j} ,$$

where the first inequality follows from the bound of $\|\widehat{g}_t^j - \widehat{g}_t^{j-1}\|$ under the event $\mathcal{E}_t(j)$ (see Equation (17)). Furthermore, by Corollary D.2, we can bound $\mathbb{E}_{t-1}\|\widehat{g}_t^0 - \nabla_t\|^2 \leq 2C^2\mathcal{V}^2\gamma$. Substituting these bounds back into Equation (20) finally gives:

$$V_t^2 \leq \mathbb{E}_{t-1} \|g_t - \nabla_t\|^2 \leq 4C^2\mathcal{V}^2\gamma + 2 \sum_{j=1}^{J_{\max}} 2^j \cdot \frac{6C^2\mathcal{V}^2\gamma}{2^j} \leq 4C^2\mathcal{V}^2\gamma + 12C^2\mathcal{V}^2\gamma J_{\max} \leq 16C^2\mathcal{V}^2\gamma \log T ,$$

where the last inequality follows from $J_{\max} \leq \log T$.

Proceeding to bound the bias, for every $t \in [T]$, we have by Jensen's inequality,

$$\|b_t\| \leq \sqrt{\mathbb{E}_{t-1} \|g_t - \nabla_t\|^2} \leq 4C\mathcal{V}\sqrt{\gamma \log T} .$$

However, for $t \in \tau_s$ the Byzantine workers are fixed, and a tighter bound can be derived. Taking expectation w.r.t J_t gives:

$$\mathbb{E}_{t-1}[g_t] = \sum_{j=1}^{\infty} 2^{-j} \cdot \mathbb{E}_{t-1} \left[\widehat{g}_t^0 + 2^j \left(\widehat{g}_t^j - \widehat{g}_t^{j-1} \right) \mathbb{1}_{\{j \leq J_{\max}\} \cap \mathcal{E}_t(j)} \right] = \mathbb{E}_{t-1}[\widehat{g}_t^0] + \sum_{j=1}^{J_{\max}} \mathbb{E}_{t-1} \left[\left(\widehat{g}_t^j - \widehat{g}_t^{j-1} \right) \mathbb{1}_{\mathcal{E}_t(j)} \right] . \tag{21}$$

Utilizing Lemma I.5, we can express each term in the sum as,

$$\mathbb{E}_{t-1} \left[\left(\widehat{g}_t^j - \widehat{g}_t^{j-1} \right) \mathbb{1}_{\mathcal{E}_t(j)} \right] = \mathbb{E}_{t-1} \left[\widehat{g}_t^j - \widehat{g}_t^{j-1} \right] - \mathbb{E}_{t-1} \left[\widehat{g}_t^j - \widehat{g}_t^{j-1} | \mathcal{E}_t(j)^c \right] \mathbb{P}_{t-1}(\mathcal{E}_t(j)^c) .$$

Denote the last term in the R.H.S by $z_t^j := \mathbb{E}_{t-1}\left[\widehat{g}_t^j - \widehat{g}_t^{j-1} | \mathcal{E}_t(j)^c\right] \mathbb{P}_{t-1}(\mathcal{E}_t(j)^c)$. Plugging this back into Equation (21), we obtain:

$$\mathbb{E}_{t-1}[g_t] = \mathbb{E}_{t-1}[\widehat{g}_t^0] + \sum_{j=1}^{J_{\max}} \left(\mathbb{E}_{t-1}[\widehat{g}_t^j - \widehat{g}_t^{j-1}] - z_t^j\right) = \mathbb{E}_{t-1}[\widehat{g}_t^{J_{\max}}] + y_t \ ,$$

where $y_t := -\sum_{j=1}^{J_{\max}} z_t^j$. Thus, by the triangle inequality and Jensen's inequality, it holds that

$$\begin{aligned} \|b_t\| = \|\mathbb{E}_{t-1}[g_t - \nabla_t]\| = \|\mathbb{E}_{t-1}[\widehat{g}_t^{J_{\max}} + y_t - \nabla_t]\| &\leq \|\mathbb{E}_{t-1}[\widehat{g}_t^{J_{\max}} - \nabla_t]\| + \|\mathbb{E}_{t-1} y_t\| \qquad (22) \\ &\leq \sqrt{\mathbb{E}_{t-1}\|\widehat{g}_t^{J_{\max}} - \nabla_t\|^2} + \mathbb{E}_{t-1}\|y_t\| \\ &\leq 2C\mathcal{V}\sqrt{\frac{\gamma}{T}} + \mathbb{E}_{t-1}\|y_t\| \ , \qquad (23) \end{aligned}$$

where the last inequality follows from Corollary D.2 and $2^{J_{\max}} \geq T/2$. Our objective now is to bound $\|y_t\|$; note that for every $j = 1, \ldots, J_{\max}$, we can bound z_t^j using Jensen's inequality as follows:

$$\|z_t^j\| = \left\|\mathbb{E}_{t-1}\left[\widehat{g}_t^j - \widehat{g}_t^{j-1} | \mathcal{E}_t(j)^c\right] \mathbb{P}_{t-1}(\mathcal{E}_t(j)^c)\right\| \leq \mathbb{E}_{t-1}\left[\|\widehat{g}_t^j - \widehat{g}_t^{j-1}\| | \mathcal{E}_t(j)^c\right] \mathbb{P}_{t-1}(\mathcal{E}_t(j)^c) \ .$$

By item 1 of Lemma D.1, for every $j = 0, \ldots, J_{\max}$ it holds that $\|\widehat{g}_t^j - \nabla_t\| \leq \mathcal{V}\sqrt{2(4\kappa_\delta + 1)}$, which implies that

$$\|\widehat{g}_t^j - \widehat{g}_t^{j-1}\| \leq \|\widehat{g}_t^j - \nabla_t\| + \|\widehat{g}_t^{j-1} - \nabla_t\| \leq 2\mathcal{V}\sqrt{2(4\kappa_\delta + 1)} \ .$$

In addition, we have by Lemma D.3 that $\mathbb{P}_{t-1}(\mathcal{E}_t(j)^c) \leq \frac{1}{2mT}$. Thus, we obtain:

$$\|z_t^j\| \leq 2\mathcal{V}\sqrt{2(4\kappa_\delta + 1)} \cdot \frac{1}{2mT} \leq \frac{2\mathcal{V}}{T}\sqrt{\frac{\gamma}{m}} \ ,$$

where we used $\frac{4\kappa_\delta + 1}{m} \leq 2\gamma$. This in turn implies, by the triangle inequality, the following bound on y_t:

$$\|y_t\| \leq \sum_{j=1}^{J_{\max}} \|z_t^j\| \leq \frac{2\mathcal{V}}{T}\sqrt{\frac{\gamma}{m}} J_{\max} \leq 2\mathcal{V}\sqrt{\frac{\gamma}{m}}\frac{\log T}{T} \ ,$$

where we used $J_{\max} \leq \log T$. Plugging this bound back into Equation (23) yields:

$$\|b_t\| \leq 2C\mathcal{V}\sqrt{\frac{\gamma}{T}} + 2\mathcal{V}\sqrt{\frac{\gamma}{m}}\frac{\log T}{T}, \quad \forall t \in \tau_s \ .$$

□

Using the established bounds on bias and variance, we derive convergence guarantees for the convex and non-convex cases.

D.1. Convex Case

The following result establishes the convergence of our approach in the convex setting.

Theorem D.5. *Assume f is convex. Under Assumptions 2.2 and 2.3, and with a (δ, κ_δ)-robust aggregator $\mathcal{A}$, consider Algorithm 2 with **Option 1** and a fixed learning rate given by*

$$\eta_t = \eta := \min\left\{\frac{D}{6C\mathcal{V}\sqrt{\gamma T \log T}}, \frac{1}{2L}\right\} \ .$$

Then, for $\overline{x}_T := \frac{1}{T}\sum_{t=2}^{T+1} x_t$, the following holds:

$$\mathbb{E}f(\overline{x}_T) - f(x^*) \leq 9CD\mathcal{V}\sqrt{\frac{\gamma \log T}{T}} + \frac{LD^2}{T} + 4CD\mathcal{V}\sqrt{\gamma \log T}\frac{|\tau_d|}{T} \ .$$

Proof. By Lemma A.1, we have

$$\mathbb{E}f(\bar{x}_T) - f(x^*) \leq \frac{D^2}{2T\eta} + \frac{\eta}{T}\sum_{t=1}^{T}\mathbb{E}V_t^2 + \frac{D}{T}\sum_{t=1}^{T}\mathbb{E}\left\|b_t\right\| .$$

Substituting the established bounds on $\|b_t\|$ and V_t^2 from Lemma D.4,

$$\begin{aligned}
\mathbb{E}f(\bar{x}_T) - f(x^*) &\leq \frac{D^2}{2T\eta} + 16\eta C^2\mathcal{V}^2\gamma\log T + \frac{D}{T}\left(\sum_{t\in\tau_d}\mathbb{E}\left\|b_t\right\| + \sum_{t\in\tau_s}\mathbb{E}\left\|b_t\right\|\right) \\
&\leq \frac{1}{2}\left(\frac{D^2}{T\eta} + 36\eta C^2\mathcal{V}^2\gamma\log T\right) + \frac{D}{T}\left(4C\mathcal{V}\sqrt{\gamma\log T}|\tau_d| + \left(2C\mathcal{V}\sqrt{\frac{\gamma}{T}} + 2\mathcal{V}\sqrt{\frac{\gamma}{m}}\frac{\log T}{T}\right)(T - |\tau_d|)\right) \\
&\leq \frac{1}{2}\left(\frac{D^2}{T\eta} + 36\eta C^2\mathcal{V}^2\gamma\log T\right) + 4CD\mathcal{V}\sqrt{\gamma\log T}\frac{|\tau_d|}{T} + 2CD\mathcal{V}\sqrt{\frac{\gamma}{T}} + 2D\mathcal{V}\sqrt{\frac{\gamma}{m}}\frac{\log T}{T} .
\end{aligned}$$

Since $\eta = \min\left\{\frac{D}{6C\mathcal{V}\sqrt{\gamma T\log T}}, \frac{1}{2L}\right\}$, applying Lemma I.7 with $a = D^2/T, b = 36C^2\mathcal{V}^2\gamma\log T$, and $c = 2L$, allows us to bound the sum of the first two terms as follows

$$\frac{D^2}{T\eta} + 36\eta C^2\mathcal{V}^2\gamma\log T \leq 12CD\mathcal{V}\sqrt{\frac{\gamma\log T}{T}} + \frac{2LD^2}{T} .$$

Plugging this bound back gives:

$$\begin{aligned}
\mathbb{E}f(\bar{x}_T) - f(x^*) &\leq 6CD\mathcal{V}\sqrt{\frac{\gamma\log T}{T}} + \frac{LD^2}{T} + 4CD\mathcal{V}\sqrt{\gamma\log T}\frac{|\tau_d|}{T} + 2CD\mathcal{V}\sqrt{\frac{\gamma}{T}} + 2D\mathcal{V}\sqrt{\frac{\gamma}{m}}\frac{\log T}{T} \\
&\leq 8CD\mathcal{V}\sqrt{\frac{\gamma\log T}{T}} + \frac{LD^2}{T} + 4CD\mathcal{V}\sqrt{\gamma\log T}\frac{|\tau_d|}{T} + CD\mathcal{V}\frac{\sqrt{\gamma\log T}}{\sqrt{m}T} \\
&\leq 9CD\mathcal{V}\sqrt{\frac{\gamma\log T}{T}} + \frac{LD^2}{T} + 4CD\mathcal{V}\sqrt{\gamma\log T}\frac{|\tau_d|}{T} ,
\end{aligned}$$

where in the second inequality we used $C = 2\sqrt{2\log\left(16m^2T\right)} \geq 2\sqrt{\log T}$ to bound the last term. □

This theorem implies the following observation.

Corollary D.6. *If $|\tau_d| \in \mathcal{O}(\sqrt{T})$, the first term dominates the convergence rate, which is $\tilde{\mathcal{O}}(D\mathcal{V}\sqrt{\gamma/T})$. Specifically, for $\kappa_\delta \in \mathcal{O}(\delta)$ this rate is given by $\tilde{\mathcal{O}}\left(D\mathcal{V}\sqrt{\left(\delta + 1/m\right)/T}\right)$.*

D.2. Non-convex Case

Having established the proof for the convex case, we move on to proving convergence in the non-convex scenario. For ease of reference, Theorem 4.1 is restated here.

Theorem 4.1. *Under Assumption 2.2, and with a (δ, κ_δ)-robust aggregator $\mathcal{A}$, consider Algorithm 2 with **Option 1** and a fixed learning rate given by*

$$\eta_t = \eta \coloneqq \min\left\{\frac{\sqrt{\Delta_1}}{3C\mathcal{V}\sqrt{L\gamma T\log T}}, \frac{1}{L}\right\} .$$

where $\gamma \coloneqq 2\kappa_\delta + \frac{1}{m}$. Then, the following holds:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla_t\|^2 \leq 12C\mathcal{V}\sqrt{\frac{L\Delta_1\gamma\log T}{T}} + \frac{2L\Delta_1 + 9C^2\mathcal{V}^2\gamma}{T} + 16C^2\mathcal{V}^2\gamma\log T\frac{|\tau_d|}{T} .$$

Proof. Utilizing Lemma A.2, we get:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\|\nabla f(x_t)\right\|^2 \leq \frac{2\Delta_1}{T\eta} + \frac{\eta L}{T}\sum_{t=1}^{T}\mathbb{E}V_t^2 + \frac{1}{T}\underbrace{\sum_{t=1}^{T}\mathbb{E}\left\|b_t\right\|^2}_{=(\star)} . \tag{24}$$

Bounding $(\star)$**:** Using the bound on the bias in item 1 of Lemma D.4, we can bound

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\left\|b_t\right\|^2 = \sum_{t\in\tau_d}\mathbb{E}\left\|b_t\right\|^2 + \sum_{t\notin\tau_d}\mathbb{E}\left\|b_t\right\|^2 &\leq 16C^2\mathcal{V}^2\gamma\log T|\tau_d| + \left(2C\mathcal{V}\sqrt{\frac{\gamma}{T}} + 2\mathcal{V}\sqrt{\frac{\gamma}{m}}\frac{\log T}{T}\right)^2(T-|\tau_d|)\\ &\leq 16C^2\mathcal{V}^2\gamma\log T|\tau_d| + 2T\left(\frac{4C^2\mathcal{V}^2\gamma}{T} + \frac{4\mathcal{V}^2\gamma\log^2 T}{mT^2}\right)\\ &\leq 16C^2\mathcal{V}^2\gamma\log T|\tau_d| + 8C^2\mathcal{V}^2\gamma + \frac{C^2\mathcal{V}^2\gamma\log T}{mT}\\ &\leq C^2\mathcal{V}^2\gamma\left(16|\tau_d|\log T + 9\right)\ ,\end{aligned}$$

where the second inequality follows from $(a+b)^2 \leq 2a^2 + 2b^2$ and $T - |\tau_d| \leq T$; the third inequality uses $8\log T \leq C^2$; and the last inequality follows from $\log T \leq mT$.

Substituting the bound on $(\star)$ and the variance bound from Lemma D.4 back into Equation (24) yields:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\|\nabla f(x_t)\right\|^2 &\leq \frac{2\Delta_1}{T\eta} + 16\eta LC^2\mathcal{V}^2\gamma\log T + \frac{C^2\mathcal{V}^2\gamma\left(16|\tau_d|\log T + 9\right)}{T}\\ &\leq 2\left(\frac{\Delta_1}{T\eta} + 9\eta LC^2\mathcal{V}^2\gamma\log T\right) + \frac{C^2\mathcal{V}^2\gamma\left(16|\tau_d|\log T + 9\right)}{T}\ .\end{aligned}$$

Utilizing Lemma I.7 with $a = \Delta_1/T, b = 9LC^2\mathcal{V}^2\gamma\log T$, and $c = L$ enables to bound the sum of the first two terms as

$$\frac{\Delta_1}{T\eta} + 9\eta LC^2\mathcal{V}^2\gamma\log T \leq 6C\mathcal{V}\sqrt{\frac{L\Delta_1\gamma\log T}{T}} + \frac{L\Delta_1}{T}\ .$$

Plugging this bound, we get:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\|\nabla f(x_t)\right\|^2 &\leq 12C\mathcal{V}\sqrt{\frac{L\Delta_1\gamma\log T}{T}} + \frac{2L\Delta_1}{T} + \frac{C^2\mathcal{V}^2\gamma\left(16|\tau_d|\log T + 9\right)}{T}\\ &= 12C\mathcal{V}\sqrt{\frac{L\Delta_1\gamma\log T}{T}} + \frac{2L\Delta_1 + 9C^2\mathcal{V}^2\gamma}{T} + 16C^2\mathcal{V}^2\gamma\frac{|\tau_d|\log T}{T}\ .\end{aligned}$$

□

E. When Worker-Momentum Fails

Next, we take a detour, to show how the worker-momentum approach may fail in the presence of Byzantine identity changes. To this end, we introduce a Byzantine identity *switching strategy*, which utilizes the momentum recursion to ensure that all workers suffer from a sufficient bias. For simplicity, we consider a setting with $m = 3$ workers[5], of which only a single worker is Byzantine in each round.

For some round t, consider the following momentum update rule with parameter $\beta \in [0, 1)$ for worker i,

$$\tilde{m}_{t,i} = \beta\tilde{m}_{t-1,i} + (1-\beta)\tilde{g}_{t,i}\ .$$

As mentioned, this update rule effectively averages the last $1/\alpha$ gradients, where $\alpha := 1-\beta$. Thus, the $1/\alpha$ rounds following a Byzantine-to-honest identity switch still heavily depend on the Byzantine behavior. Intuitively, this 'healing phase' of $1/\alpha$ rounds is the time required for the worker to produce informative honest updates. Our attack leverages this property to perform an identity switch once in every $1/3\alpha$, to maintain all workers under the Byzantine effect, i.e., to prevent workers from completely 'healing' from the attack. Recall that existing approaches to Byzantine-resilient strategy suggests choosing $\alpha \approx 1/\sqrt{T}$ to establish theoretical guarantees (see, e.g., Karimireddy et al., 2021; Allouah et al., 2023). Thus, henceforth, we will assume $\alpha \leq 1/6$ and, for the sake of simplicity, that $1/3\alpha$ is an integer.

[5]For general m, we can divide the workers into 3 groups and apply our switching strategy to these groups.

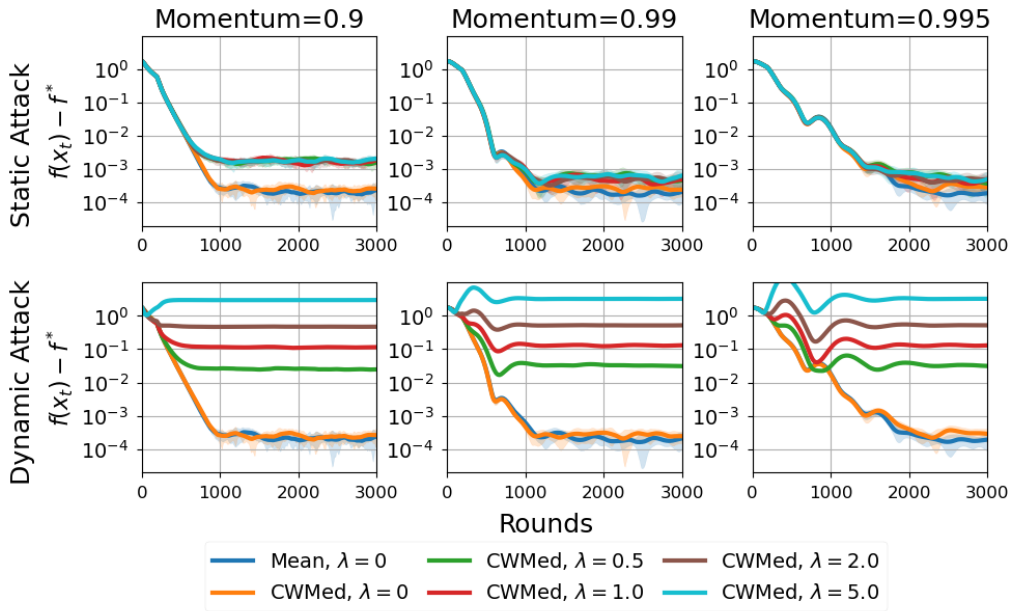

Figure 3. Optimality gap ($f(x_t) - f^*$) under static (**top**) and dynamic (**bottom**) attacks across various momentum parameters and for different attack strengths ($\lambda = 0, 0.5, 1, 2, 5$). The average and 95% confidence interval are presented over 20 random seeds.

We divide the T training rounds into epochs (i.e., windows) of size $1/\alpha \approx \sqrt{T}$. Within these epochs, we perform an identity switch once in every $1/3\alpha$ rounds, periodically, implying 3 identity switches per-epoch and $3\alpha T \in \mathcal{O}(\sqrt{T})$ switches overall. Concretely, denoting by $\tilde{g}_{t,i}$ the gradient used by worker $i \in \{1, 2, 3\}$ at time t to perform momentum update, we consider the following attack strategy:

$$\tilde{g}_{t,i} = g_{t,i} + v_t \cdot \mathbb{1}\left\{t \mod \frac{1}{\alpha} \in \left[\frac{i-1}{3\alpha} + 1, \frac{i}{3\alpha}\right]\right\} ,$$

where $g_{t,i}$ is an honest stochastic gradient and v_t is an attack vector to be defined later. Note that under this attack strategy, there is indeed only a single Byzantine machine at a time in all rounds.

Denote by $\tilde{m}_{t,i} := m_{t,i} + b_{t,i}$ the momentum used by worker i in round t, where $m_{t,i} = (1-\alpha)m_{t-1,i} + \alpha g_{t,i}$ is the honest momentum (without Byzantine attack) and $b_{t,i}$ is the bias introduced by our attack. We want to find a recursion for $b_{t,i}$ to characterize the dynamics of the deviation from the honest momentum protocol. By plugging $\tilde{m}_{t-1,i}$ and $\tilde{g}_{t,i}$, we get:

$$\begin{aligned}\tilde{m}_{t,i} &= (1-\alpha)\left(m_{t-1,i} + b_{t-1,i}\right) + \alpha\left(g_{t,i} + v_t \cdot \mathbb{1}\left\{t \mod \frac{1}{\alpha} \in \left[\frac{i-1}{3\alpha} + 1, \frac{i}{3\alpha}\right]\right\}\right) \\ &= m_{t,i} + (1-\alpha)b_{t-1,i} + \alpha v_t \cdot \mathbb{1}\left\{t \mod \frac{1}{\alpha} \in \left[\frac{i-1}{3\alpha} + 1, \frac{i}{3\alpha}\right]\right\} .\end{aligned}$$

This implies the following recursion for the bias of worker i,

$$b_{t,i} = (1-\alpha)b_{t-1,i} + \alpha v_t \cdot \mathbb{1}\left\{t \mod \frac{1}{\alpha} \in \left[\frac{i-1}{3\alpha} + 1, \frac{i}{3\alpha}\right]\right\} .$$

Let $v \in \mathbb{R}^d$ be some fixed vector. By carefully choosing v_t, as we describe next, we ensure that $b_{t,i} = v$ for all rounds t under which worker i is Byzantine. Note that $b_{1,1} = v_1$, and $b_{1,2} = b_{1,3} = 0$. We distinguish between the first epoch and the following ones. For the first epoch, i.e., $t \in [1/\alpha]$, we choose:

$$v_t = v \cdot \begin{cases} 1/\alpha, & t \in \left\{\frac{1}{3\alpha} + 1, \frac{2}{3\alpha} + 1\right\} \\ 1, & \text{otherwise.} \end{cases}$$

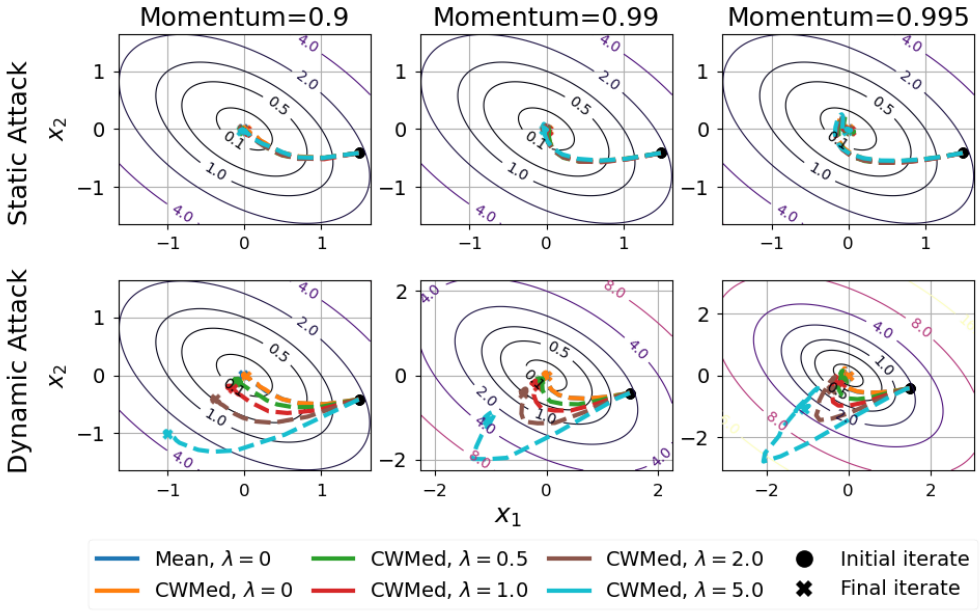

Figure 4. Optimization trajectories under static (**top**) and dynamic (**bottom**) attacks for a range of momentum parameters and for different attack strengths ($\lambda = 0, 0.5, 1, 2, 5$). Note that under the dynamic attack, the algorithm converges to a sub-optimal solution.

For the subsequent epochs, i.e., $t \in \left[\frac{1}{\alpha} + 1, T\right]$, we choose:

$$v_t = v \cdot \begin{cases} \left(1 - (1-\alpha)^{2/3\alpha}\right)/\alpha, & t \mod \frac{1}{\alpha} = 1 \\ 1, & \text{otherwise.} \end{cases}$$

The following lemma establishes that all worker momentums are sufficiently biased under this attack.

Lemma E.1. *Given the above attack, for any t starting from the second epoch, the following holds for $i = 1, 2, 3$:*

$$\tilde{m}_{t,i} = m_{t,i} + \theta_{t,i} v \ ,$$

where $\theta_{t,i} \in [\theta_{\min}, 1]$ for each i and for all $t > 1/\alpha$. Here, $\theta_{\min} := \left(\frac{5}{6}\right)^4 > 0.48$.

Proof. Let us examine the bias dynamics of the first worker under the described attack.

In the first third of each epoch, namely, for $t \bmod \frac{1}{\alpha} \in \left[\frac{1}{3\alpha} + 1, \frac{1}{3\alpha}\right]$, we have a fixed bias of $b_{t,1} = v$. For the remaining two thirds, we have an exponential bias decay, where $b_{t,1} = \theta_{t,1} v$ for $\theta_{t,1} := (1-\alpha)^{\left(t - \frac{1}{3\alpha}\right) \bmod \frac{1}{\alpha}}$. Since β^t is decreasing with t for $\beta < 1$, the coefficients sequence $\theta_{t,i}$ obtains its minimum at the end of each epoch, when $t \bmod \frac{1}{\alpha} = 0$, given by $(1-\alpha)^{2/3\alpha}$. For $\alpha \leq 1/6$, it holds that $(1-\alpha)^{2/3\alpha} \geq \left(\frac{5}{6}\right)^4 := \theta_{\min}$. Since $\theta_{t,2}$ and $\theta_{t,3}$ are simply a shift of $\theta_{t,1}$ by $1/3\alpha$ and $2/3\alpha$ rounds, respectively, they satisfy the same bounds. □

Given the above bias statement, any robust aggregation rule has no ability to infer an unbiased momentum path, and it would arbitrarily fail as v is unbounded.

We provide an empirical evidence to demonstrate our observation using a simple 2D quadratic example. Consider the function $f(x) = \frac{1}{2} x^T A x$ with $x \in \mathbb{R}^2$ and A is the matrix $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. In our attack setup, each worker ($i = 1, 2, 3$) employs momentum-SGD. The honest gradient oracle for each worker is defined as $g_{t,i} = \nabla_t + n_{t,i}$, where $n_{t,i} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2 I)$ with $\sigma = 0.5$. We set the attack vector to $v = \lambda \cdot \begin{bmatrix} 1 & 1 \end{bmatrix}^\top$, and examine various values of $\lambda \in \{0, 0.5, 1, 2, 5\}$ ($\lambda = 0$ corresponds to the Byzantine-free setting). At the server level, we process the worker-momentums using either simple averaging (Mean) or coordinate-wise median (CWMed). The aggregated momentum $\widehat{m}_t$ is then used in the update rule:

$x_{t+1} = x_t - \eta \widehat{m}_t$ with a learning rate $\eta = 5 \cdot 10^{-3}$ over $T = 3000$ rounds. We experiment with various momentum parameters β from $\{0.9, 0.99, 0.995\}$, corresponding to α values of $\{0.1.0.01, 0.005\}$, and repeat each experiment with 20 different random seeds. Note that these values correspond to $|\tau_d| = 999, 90, 45$ rounds with Byzantine identity changes. Additionally, we include a 'static attack' scenario where only the first worker is consistently Byzantine, using $\tilde{g}_{t,1} = g_{t,1} + v$ throughout all rounds.

In Figure 3, we illustrate the optimality gap during the training process. Notably, in the presence of a dynamic attack (where $\lambda > 0$), we observe that the error plateaus at a sub-optimal level for all values of the momentum parameter. Furthermore, there is a clear trend that shows an increase in the final error magnitude in direct proportion to the strength of the attack (as λ increases). This trend is distinct from what we observe under a static attack, where such a correlation between the attack strength and final error is not apparent.

Correspondingly, in Figure 4, we present a representative example showcasing the optimization paths under the influence of the static and dynamic attacks, with various momentum parameters. The trajectories visibly diverge towards sub-optimal points under dynamic attacks, with the divergence growing as the attack strength, i.e., λ, is increased. In contrast, the static attack scenarios reveal paths that remain relatively stable despite changes in attack strength. This visual illustration underscores the possible failure of the worker-momentum approach under dynamic Byzantine attack.

F. Properties of the MFM Aggregator

In this section, we establish the properties of the MFM aggregator, crucial for our analysis of Section 5. Additionally, in Appendix F.1, we demonstrate that MFM does not meet the (δ, κ)-robustness criteria.

We assume the gradient noise is bounded (Assumption 2.2) and consider the MFM aggregator with inputs $\bar{g}_1^N, \ldots, \bar{g}_m^N$ as in Lemma 3.3 and threshold parameter set to $\mathcal{T}_p^N = 2C_p\mathcal{V}/\sqrt{N}$, where $C_p := \sqrt{8\log(2m/p)}$.

Initially, we introduce the following event, under which we derive valuable insights regarding Algorithm 3.

$$\mathcal{B} = \left\{ \forall i \in \mathcal{G} : \left\| \bar{g}_i^N - \nabla f(x) \right\| \leq \frac{\mathcal{T}_p^N}{4} = \mathcal{V}\sqrt{\frac{2\log\left(2m/p\right)}{N}} \right\} . \tag{25}$$

Here, we analyze Algorithm 3 assuming the honest workers are fixed when computing mini-batches.

First, we show that $\mathcal{B}$ is satisfied with high probability.

Lemma F.1. *For every $x \in \mathbb{R}$, it holds that $\mathbb{P}(\mathcal{B}) \geq 1 - p$.*

Proof. Since for every $i \in \mathcal{G}$, $\bar{g}_i^N = \frac{1}{N}\sum_{n=1}^N \nabla F(x; \xi_i^n)$, where $\|\nabla F(x; \xi_i^n) - \nabla f(x)\| \leq \mathcal{V}$, utilizing Lemma I.1 implies that with probability at least $1 - \tilde{p}$, we have $\left\|\bar{g}_i^N - \nabla f(x)\right\| \leq \mathcal{V}\sqrt{\frac{2\log(2/\tilde{p})}{N}}$. Employing the union bound and using $|\mathcal{G}| \leq m$ establishes the result. □

Next, we establish some results assuming $\mathcal{B}$ holds.

Lemma F.2. *Under the event $\mathcal{B}$, the following holds:*

1. *The set $\mathcal{M}$ is not empty.*
2. *$\mathcal{G} \subseteq \widehat{\mathcal{G}}$.*
3. *For every $i \in \widehat{\mathcal{G}}$, we have $\left\|\bar{g}_i^N - \nabla f(x)\right\| \leq 2\mathcal{T}_p^N$.*

Proof. Following the proof of Claim 3.4 in Alistarh et al. (2018): under the event $\mathcal{B}$, for every $i, j \in \mathcal{G}$, we have by the triangle inequality that $\left\|\bar{g}_i^N - \bar{g}_j^N\right\| \leq \left\|\bar{g}_i^N - \nabla f(x)\right\| + \left\|\bar{g}_j^N - \nabla f(x)\right\| \leq \mathcal{T}_p^N/2$. Since $\delta < 1/2$, every $i \in \mathcal{G}$ is also in $\mathcal{M}$, namely, $\mathcal{G} \subseteq \mathcal{M}$, which concludes the first part.

For the second and third parts, we first show that $\|g_{\text{med}} - \nabla f(x)\| \leq 3\mathcal{T}_p^N/4$; assuming $\|g_{\text{med}} - \nabla f(x)\| > 3\mathcal{T}_p^N/4$, we get by the triangle inequality that $\left\|g_{\text{med}} - \bar{g}_i^N\right\| > \mathcal{T}_p^N/2$ for every $i \in \mathcal{G}$, thus contradicting the definition of g_{med} (chosen from $\mathcal{M}$) as $|\mathcal{G}| > m/2$.

To prove the second part, fix some $i \in \mathcal{G}$. By the triangle inequality: $\left\|\bar{g}_i^N - g_{\text{med}}\right\| \leq \left\|\bar{g}_i^N - \nabla f(x)\right\| + \left\|g_{\text{med}} - \nabla f(x)\right\| \leq \mathcal{T}_p^N/4 + 3\mathcal{T}_p^N/4 = \mathcal{T}_p^N$, where the bound on $\left\|\bar{g}_i^N - \nabla f(x)\right\|$ follows from the definition of $\mathcal{B}$. This bound implies that $i \in \widehat{\mathcal{G}}$ as well, concluding the second part.

Finally, note that for any $i \in \widehat{\mathcal{G}}$, the triangle inequality implies $\left\|\bar{g}_i^N - \nabla f(x)\right\| \leq \left\|\bar{g}_i^N - g_{\text{med}}\right\| + \left\|g_{\text{med}} - \nabla f(x)\right\| \leq \mathcal{T}_p^N + 3\mathcal{T}_p^N/4 \leq 2\mathcal{T}_p^N$, where $\left\|\bar{g}_i^N - g_{\text{med}}\right\| \leq \mathcal{T}_p^N$ holds for any $i \in \widehat{\mathcal{G}}$ as established in the previous part. □

With these insights, we now prove Lemma F.3, which, similarly to Lemma D.1, provides deterministic and high-probability bounds on the aggregation error.

Lemma F.3. *Consider the setting in Lemma D.1 and let $\widehat{g}^N \leftarrow$ MFM$(\bar{g}_1^N, \ldots, \bar{g}_m^N; \mathcal{T}_p^N)$ be the output of Algorithm 3 with $\mathcal{T}_p^N := 2C_p\mathcal{V}/\sqrt{N}$, where $C_p := \sqrt{8\log(2m/p)}$. Then,*

1. $\left\|\widehat{g}^N - \nabla f(x)\right\|^2 \leq \frac{9}{2}(\mathcal{T}_p^N)^2 + 2\mathcal{V}^2 + \left\|\nabla f(x)\right\|^2 \leq 20C_p^2\mathcal{V}^2 + \left\|\nabla f(x)\right\|^2$.
2. *With probability at least $1-p$,*

$$\left\|\widehat{g}^N - \nabla f(x)\right\|^2 \leq \frac{2C_p^2\mathcal{V}^2}{N}\left(\frac{1}{m} + 128\delta^2\right) .$$

Proof. Denote: $\nabla := \nabla f(x)$. To prove the first part, we consider two cases: when $\mathcal{M}$ is either empty or non-empty. If $\mathcal{M} = \emptyset$, then $\widehat{g}^N = 0$, leading to $\left\|\widehat{g}^N - \nabla\right\|^2 = \left\|\nabla\right\|^2$. In the case where $\mathcal{M}$ is non-empty, the triangle inequality implies that

$$\left\|\widehat{g}^N - \nabla\right\| = \left\|\frac{1}{|\widehat{\mathcal{G}}|}\sum_{i\in\widehat{\mathcal{G}}}\left(\bar{g}_i^N - \nabla\right)\right\| \leq \frac{1}{|\widehat{\mathcal{G}}|}\sum_{i\in\widehat{\mathcal{G}}}\left\|g_i^N - \nabla\right\| .$$

Since $g_{\text{med}} = \bar{g}_i^N$ for some $i \in \mathcal{M}$, by the definition of $\mathcal{M}$, there are more than $m/2$ machines whose distance from g_{med} is bounded by $\mathcal{T}_p^N/2$. Since there are at most $\delta m < m/2$ Byzantine workers, it implies that at least one of the above workers is good; denote one such worker by $\ell \in \mathcal{G}$. Thus, for every $i \in \widehat{\mathcal{G}}$ we can bound,

$$\left\|\bar{g}_i^N - \nabla\right\| \leq \underbrace{\left\|\bar{g}_i^N - g_{\text{med}}\right\|}_{\leq \mathcal{T}_p^N} + \underbrace{\left\|g_{\text{med}} - \bar{g}_\ell^N\right\|}_{\leq \mathcal{T}_p^N/2} + \underbrace{\left\|\bar{g}_\ell^N - \nabla\right\|}_{\leq \mathcal{V}} \leq \frac{3\mathcal{T}_p^N}{2} + \mathcal{V} ,$$

where the first bound is by the definition of $\widehat{\mathcal{G}}$, the second bound is a consequence of the chosen ℓ (and the definition of g_{med}), and the final bound is trivial for any good worker. Since this bound holds for any $i \in \widehat{\mathcal{G}}$, it also holds for $\widehat{g}^N$. Thus, using $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$\left\|\widehat{g}^N - \nabla\right\|^2 \leq 2\left(\frac{9}{4}(\mathcal{T}_p^N)^2 + \mathcal{V}^2\right) = \frac{9}{2}(\mathcal{T}_p^N)^2 + 2\mathcal{V}^2 .$$

Overall, in any case, we have that

$$\begin{aligned}\left\|\widehat{g}^N - \nabla\right\|^2 \leq \max\left\{\left\|\nabla\right\|^2, \frac{9}{2}(\mathcal{T}_p^N)^2 + 2\mathcal{V}^2\right\} &\leq \frac{9}{2}(\mathcal{T}_p^N)^2 + 2\mathcal{V}^2 + \left\|\nabla\right\|^2 \\ &= \frac{9}{2}\cdot\frac{4C_p^2\mathcal{V}^2}{N} + 2\mathcal{V}^2 + \left\|\nabla\right\|^2 \\ &\leq 20C_p^2\mathcal{V}^2 + \left\|\nabla\right\|^2 ,\end{aligned}$$

where we used $C_p^2 \geq 1$. This concludes the first part.

For the second part, we denote the average of honest workers by $\bar{g}^N := \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\bar{g}_i^N$ and define the following events:

$$\widehat{\mathcal{C}} := \left\{\left\|\widehat{g}^N - \nabla\right\| \leq 2\mathcal{V}\sqrt{\frac{\log(2/p)}{mN}} + 4\delta\mathcal{T}_p^N\right\},$$

$$\bar{\mathcal{C}} := \left\{\left\|\bar{g}^N - \nabla\right\| \leq 2\mathcal{V}\sqrt{\frac{\log(2/p)}{mN}}\right\} .$$

Our objective is to establish that $\widehat{\mathcal{C}}$ holds with probability at least $1-2p$. Recall that under the event $\mathcal{B}$ (Equation (25)), the set $\mathcal{M}$ is non-empty (item 1 of Lemma F.2), implying $\widehat{g}^N = \frac{1}{|\widehat{\mathcal{G}}|}\sum_{i\in\widehat{\mathcal{G}}}\bar{g}_i^N$, and $\mathcal{G}\subseteq\widehat{\mathcal{G}}$. Hence, under the event $\mathcal{B}$, we have that

$$\begin{aligned}\left\|\widehat{g}^N-\nabla\right\| = \left\|\frac{1}{|\widehat{\mathcal{G}}|}\sum_{i\in\widehat{\mathcal{G}}}(\bar{g}_i^N-\nabla)\right\| &= \left\|\frac{1}{|\widehat{\mathcal{G}}|}\sum_{i\in\mathcal{G}}(\bar{g}_i^N-\nabla)+\frac{1}{|\widehat{\mathcal{G}}|}\sum_{i\in\widehat{\mathcal{G}}\setminus\mathcal{G}}(\bar{g}_i^N-\nabla)\right\| \\ &\le \frac{|\mathcal{G}|}{|\widehat{\mathcal{G}}|}\left\|\bar{g}^N-\nabla\right\|+\frac{1}{|\widehat{\mathcal{G}}|}\sum_{i\in\widehat{\mathcal{G}}\setminus\mathcal{G}}\left\|\bar{g}_i^N-\nabla\right\| \\ &\le \left\|\bar{g}^N-\nabla\right\|+\frac{2}{m}|\widehat{\mathcal{G}}\setminus\mathcal{G}|\cdot 2\mathcal{T}_p^N \\ &\le \left\|\bar{g}^N-\nabla\right\|+4\delta\mathcal{T}_p^N\ , \end{aligned} \tag{26}$$

where the first inequality follows from the triangle inequality and the fact that $\sum_{i\in\mathcal{G}}(\bar{g}_i^N-\nabla)=|\mathcal{G}|(\bar{g}^N-\nabla)$; the second inequality is due to $|\widehat{\mathcal{G}}|\ge|\mathcal{G}|\ge m/2$ and item 3 of Lemma F.2, namely, $\|\bar{g}_i^N-\nabla\|\le 2\mathcal{T}_p^N$ for all $i\in\widehat{\mathcal{G}}$; and the last inequality follows from $|\widehat{\mathcal{G}}\setminus\mathcal{G}|\le|[m]\setminus\mathcal{G}|=\delta m$.

Based on Equation (26), we infer that, under the event $\mathcal{B}$, if $\bar{\mathcal{C}}$ occurs, then so does $\widehat{\mathcal{C}}$, implying $\mathbb{P}(\widehat{\mathcal{C}}|\mathcal{B})\ge\mathbb{P}(\bar{\mathcal{C}}|\mathcal{B})$. Furthermore, we have by Lemma F.1 that $\mathbb{P}(\mathcal{B})\ge 1-p>0$. Combining these properties and utilizing Lemma I.6 yields:

$$\mathbb{P}(\widehat{\mathcal{C}})\ge\mathbb{P}(\bar{\mathcal{C}})-\mathbb{P}(\mathcal{B}^c)\ge\mathbb{P}(\bar{\mathcal{C}})-p\ . \tag{27}$$

By Lemma I.1, it holds with probability at least $1-p$ that

$$\left\|\bar{g}^N-\nabla\right\| = \frac{1}{|\mathcal{G}|N}\left\|\sum_{i\in\mathcal{G}}\sum_{n\in[N]}(\nabla F(x;\xi_i^n)-\nabla)\right\| \le \mathcal{V}\sqrt{\frac{2\log(2/p)}{|\mathcal{G}|N}}\le 2\mathcal{V}\sqrt{\frac{\log(2/p)}{mN}}\ .$$

In other words, $\mathbb{P}(\bar{\mathcal{C}})\ge 1-p$, indicating, as per Equation (27), that with a probability of at least $1-2p$,

$$\left\|\widehat{g}^N-\nabla\right\|\le 2\mathcal{V}\sqrt{\frac{\log(2/p)}{mN}}+4\delta\mathcal{T}_p^N = 2\mathcal{V}\sqrt{\frac{\log(2/p)}{mN}}+16\delta\mathcal{V}\sqrt{\frac{2\log(2m/p)}{N}} = 2\mathcal{V}\sqrt{\frac{\log(2m/p)}{N}}\left(\frac{1}{\sqrt{m}}+8\sqrt{2}\delta\right)\ .$$

This result finally implies that with probability at least $1-p$,

$$\left\|\widehat{g}^N-\nabla\right\|^2\le\frac{4\mathcal{V}^2\log(4m/p)}{N}\left(\frac{1}{\sqrt{m}}+8\sqrt{2}\delta\right)^2\le\frac{8\mathcal{V}^2\log(4m/p)}{N}\left(\frac{1}{m}+128\delta^2\right)\le\frac{2C_p^2\mathcal{V}^2}{N}\left(\frac{1}{m}+128\delta^2\right)\ ,$$

where we used $(a+b)^2\le 2(a^2+b^2)$ and $8\log(4m/p)\le 16\log(4m/p)=2C_p^2$. □

Combining the bounds established in Lemma F.3, we can derive an upper bound on the expected (squared) aggregation error, mirroring Corollary D.2.

Corollary F.4 (MSE of Aggregated Gradients)**.** *Consider the setting in Lemma D.1, let $\widehat{g}^N\leftarrow$ MFM$(\bar{g}_1^N,\dots,\bar{g}_m^N;\mathcal{T}^N)$ be the output of Algorithm 3 with $\mathcal{T}^N:=2C\mathcal{V}/\sqrt{N}=4\mathcal{V}\sqrt{2\log(16m^2T)/N}$, and assume that $N\le T$. Then,*

$$\mathbb{E}\left\|\widehat{g}^N-\nabla f(x)\right\|^2\le\frac{\tilde{C}^2\mathcal{V}^2\tilde{\gamma}}{N}+\frac{\|\nabla f(x)\|^2}{8mT}\ ,$$

where $\tilde{C}^2:=8C^2=64\log\big(16m^2T\big)$ and $\tilde{\gamma}:=\frac{1}{m}+32\delta^2$.

Proof. Utilizing the previous lemma, we have with probability at least $1-\frac{1}{8mT}$,

$$\left\|\widehat{g}^N-\nabla f(x)\right\|^2\le\frac{2C^2\mathcal{V}^2}{N}\left(\frac{1}{m}+128\delta^2\right)\ .$$

Additionally, we always have that

$$\left\|\widehat{g}^N - \nabla f(x)\right\|^2 \leq 20C^2\mathcal{V}^2 + \|\nabla f(x)\|^2 \ .$$

Thus, by the law of total expectation, we get

$$\begin{aligned}
\mathbb{E}\left\|\widehat{g}^N - \nabla f(x)\right\|^2 &\leq \frac{2C^2\mathcal{V}^2}{N}\left(\frac{1}{m} + 128\delta^2\right) + \left(20C^2\mathcal{V}^2 + \|\nabla f(x)\|^2\right) \cdot \frac{1}{8mT} \\
&= \frac{2C^2\mathcal{V}^2}{N}\left(\frac{1}{m} + 128\delta^2\right) + \frac{5C^2\mathcal{V}^2}{2mT} + \frac{\|\nabla f(x)\|^2}{8mT} \\
&= \frac{C^2\mathcal{V}^2}{N}\left(\frac{9}{2m} + 256\delta^2\right) + \frac{\|\nabla f(x)\|^2}{8mT} \\
&\leq \frac{8C^2\mathcal{V}^2}{N}\left(\frac{1}{m} + 32\delta^2\right) + \frac{\|\nabla f(x)\|^2}{8mT} \\
&= \frac{\tilde{C}^2\mathcal{V}^2\tilde{\gamma}}{N} + \frac{\|\nabla f(x)\|^2}{8mT} \ ,
\end{aligned}$$

where the second inequality follows from the assumption that $N \leq T$. □

The bound in Corollary F.4 closely resembles that in Corollary D.2, differing by an additional factor of $\frac{\|\nabla f(x)\|^2}{8mT}$. This variation stems from the specific structure of the MFM aggregator, specifically due to the rare event where the set $\mathcal{M}$ is empty, leading to an output of $\widehat{g} = 0$.

F.1. MFM is not (δ, κ)-robust

Consider Algorithm 3 with a threshold $\mathcal{T} > 0$, and let $x \in \mathbb{R}^d$. Suppose every honest worker $i \in \mathcal{G}$ provides the true gradient, $g_i = \nabla f(x)$, while every Byzantine worker submits $g_i = \nabla f(x) + \frac{3}{4}\mathcal{T}v$, for some $v \in \mathbb{R}^d$ with $\|v\| = 1$. In this case, $\mathcal{M}$ is not empty as $\mathcal{M} = \mathcal{G}$. Since all vectors are within $\frac{3}{4}\mathcal{T}$ of each other, $\widehat{\mathcal{G}}$ contains all workers. Consequently, the aggregated gradient, $\widehat{g}$, is given by $\nabla f(x) + \frac{3}{4}\mathcal{T}\delta v$. Denoting the average of honest workers by $\bar{g} = \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} g_i = \nabla f(x)$, the above implies a nonzero aggregation error $\|\widehat{g} - \bar{g}\|$, while the 'variance' among honest workers, $\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} \|g_i - \bar{g}\|^2$, remains zero. This scenario fails to satisfy Definition 3.2.

G. AdaGrad with Biased Gradients

Consider the AdaGrad-Norm (Levy, 2017; Ward et al., 2020; Faw et al., 2022) (also known as AdaSGD, Attia & Koren 2023) update rule, defined for some parameter $\eta_0 > 0$ as follows:

$$x_{t+1} = \Pi_{\mathcal{K}}(x_t - \eta_t g_t), \quad \eta_t = \frac{\eta_0}{\sqrt{\sum_{s=1}^t \|g_s\|^2}} \ , \tag{AdaGrad-Norm}$$

where g_t has bias $b_t := \mathbb{E}[g_t - \nabla f(x_t)|x_t]$ and variance $V_t^2 := \mathbb{E}[\|g_t - \mathbb{E}g_t\|^2 |x_t]$.

In Appendices G.1 and G.2, convergence bounds for (AdaGrad-Norm) are deduced for convex and non-convex objectives, respectively.

G.1. Convex Analysis

We commence with a lemma that establishes a second-order bound on the linearized regret of (AdaGrad-Norm), essential in our convex analysis. The proof is included for completeness.

Lemma G.1 (Levy, 2017, Theorem 1.1)**.** *Suppose Assumption 2.3 holds, i.e., the domain $\mathcal{K}$ is bounded with diameter $D := \max_{x,y\in\mathcal{K}} \|x - y\|$ and consider* (AdaGrad-Norm)*. Then, for every $u \in \mathcal{K}$, the iterates $x_1, \ldots, x_T$ satisfy:*

$$\sum_{t=1}^T g_t^\top (x_t - u) \leq \left(\frac{D^2}{2\eta_0} + \eta_0\right)\sqrt{\sum_{t=1}^T \|g_t\|^2} \ .$$

Proof. For every $u \in \mathcal{K}$, we have that

$$\|x_{t+1} - u\|^2 \leq \|x_t - u\|^2 - 2\eta_t g_t^\top (x_t - u) + \eta^2 \|g_t\|^2 \ .$$

Rearranging terms, we get:

$$g_t^\top (x_t - u) \leq \frac{\|x_t - u\|^2 - \|x_{t+1} - u\|^2}{2\eta_t} + \frac{\eta_t}{2} \|g_t\|^2 \ .$$

Summing over $t \in [T]$, we then obtain:

$$\begin{aligned}
\sum_{t=1}^{T} g_t^\top (x_t - u) &\leq \frac{\|x_1 - u\|^2}{2\eta_1} + \sum_{t=2}^{T} \frac{\|x_t - u\|^2}{2} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) + \frac{1}{2} \sum_{t=1}^{T} \eta_t \|g_t\|^2 \\
&\leq \frac{D^2}{2} \left(\frac{1}{\eta_1} + \sum_{t=2}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \right) + \frac{\eta_0}{2} \sum_{t=1}^{T} \frac{\|g_t\|^2}{\sqrt{\sum_{s=1}^{t} \|g_s\|^2}} \\
&= \frac{D^2}{2\eta_T} + \frac{\eta_0}{2} \sum_{t=1}^{T} \frac{\|g_t\|^2}{\sqrt{\sum_{s=1}^{t} \|g_s\|^2}} \\
&\leq \frac{D^2}{2\eta_0} \sqrt{\sum_{t=1}^{T} \|g_t\|^2} + \eta_0 \sqrt{\sum_{t=1}^{T} \|g_t\|^2} \\
&= \left(\frac{D^2}{2\eta_0} + \eta_0 \right) \sqrt{\sum_{t=1}^{T} \|g_t\|^2} \ ,
\end{aligned}$$

where the second inequality uses $\|x_t - u\|^2 \leq D^2$ for every $t \in [T]$ and $\eta_t \leq \eta_{t-1}$, and the final inequality stems from Lemma I.4. □

We now establish a regret bound for AdaGrad-Norm.

Lemma G.2. *Assume f is convex. Under Assumption 2.3, the iterates of* (AdaGrad-Norm) *satisfy:*

$$\mathbb{E}\mathcal{R}_T \leq D \sqrt{2 \sum_{t \in [T]} \mathbb{E} V_t^2} + 2D \sqrt{\sum_{t \in [T]} \mathbb{E}\|b_t\|^2} + D \sum_{t \in [T]} \mathbb{E}\|b_t\| + 2D \sqrt{\sum_{t \in [T]} \mathbb{E}\|\nabla f(x_t)\|^2} \ ,$$

where $\mathcal{R}_T \coloneqq \sum_{t \in [T]} (f(x_t) - f(x^*))$.

Proof. By the convexity of f,

$$\mathbb{E}\mathcal{R}_T \leq \sum_{t \in [T]} \mathbb{E}[\nabla f(x_t)^\top (x_t - x^*)] = \underbrace{\mathbb{E}\left[\sum_{t \in [T]} g_t^\top (x_t - x^*) \right]}_{=(A)} + \underbrace{\sum_{t \in [T]} \mathbb{E}[-b_t^\top (x_t - x^*)]}_{=(B)} \ . \tag{28}$$

Bounding (A). Utilizing Lemma G.1 with $\eta_0 = D/\sqrt{2}$ and Jensen's inequality, we obtain:

$$\mathbb{E}\left[\sum_{t \in [T]} g_t^\top (x_t - x^*) \right] \leq D\mathbb{E}\left[\sqrt{2 \sum_{t \in [T]} \|g_t\|^2} \right] \leq D \sqrt{2 \sum_{t \in [T]} \mathbb{E}\|g_t\|^2} \ .$$

We can bound the second moment of g_t as follows:

$$\mathbb{E}_{t-1}\|g_t\|^2 = \mathbb{E}_{t-1}\|g_t - \mathbb{E}_{t-1} g_t\|^2 + \|\mathbb{E}_{t-1} g_t\|^2 \leq V_t^2 + 2\|b_t\|^2 + 2\|\nabla f(x_t)\|^2 \ , \tag{29}$$

where we used $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. Plugging this bound back, we get that (A) is bounded as

$$\mathbb{E}\left[\sum_{t\in[T]} g_t^\top (x_t - x^*)\right] \leq D\sqrt{2\sum_{t\in[T]} \mathbb{E}V_t^2} + 2D\sqrt{\sum_{t\in[T]} \mathbb{E}\|b_t\|^2} + 2D\sqrt{\sum_{t\in[T]} \mathbb{E}\|\nabla f(x_t)\|^2} \ .$$

Bounding (B)**.** By Cauchy-Schwarz inequality and Assumption 2.3,

$$\sum_{t\in[T]} \mathbb{E}[-b_t^\top (x_t - x^*)] \leq \sum_{t\in[T]} \mathbb{E}[\|b_t\| \|x_t - x^*\|] \leq D\sum_{t\in[T]} \mathbb{E}\|b_t\| \ .$$

Incorporating the bounds on (A) and (B) into Equation (28) concludes the proof, as

$$\mathbb{E}\mathcal{R}_T \leq D\sqrt{2\sum_{t\in[T]} \mathbb{E}V_t^2} + 2D\sqrt{\sum_{t\in[T]} \mathbb{E}\|b_t\|^2} + D\sum_{t\in[T]} \mathbb{E}\|b_t\| + 2D\sqrt{\sum_{t\in[T]} \mathbb{E}\|\nabla f(x_t)\|^2} \ .$$

□

G.2. Non-Convex Analysis

The subsequent lemma establishes an upper bound on the sum of squared gradient norms when utilizing AdaGrad-Norm for bounded functions.

Lemma G.3. *Assume f is bounded by M, i.e., $\max_x |f(x)| \leq M$, and consider* (AdaGrad-Norm) *with $\mathcal{K} = \mathbb{R}^d$. The iterates $x_1, \ldots, x_T$ satisfy:*

$$\sum_{t=1}^T \|\nabla f(x_t)\|^2 \leq \left(\frac{2M}{\eta_0} + \eta_0 L\right)\sqrt{\sum_{t=1}^T \|g_t\|^2} + \sum_{t=1}^T (\nabla f(x_t) - g_t)^\top \nabla f(x_t) \ .$$

Proof. By the smoothness of f, for every $x, y \in \mathbb{R}^d$ we have that $f(y) \leq f(x) + \nabla f(x)^\top (y-x) + \frac{L}{2}\|y-x\|^2$. Plugging-in the update rule $x_{t+1} = x_t - \eta_t g_t$, we obtain:

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) - \eta_t \nabla f(x_t)^\top g_t + \frac{L\eta_t^2}{2}\|g_t\|^2 \\ &= f(x_t) - \eta_t \|\nabla f(x_t)\|^2 + \eta_t (\nabla f(x_t) - g_t)^\top \nabla f(x_t) + \frac{L\eta_t^2}{2}\|g_t\|^2 \ . \end{aligned}$$

Rearranging terms and dividing by η_t gives:

$$\|\nabla f(x_t)\|^2 \leq \frac{f(x_t) - f(x_{t+1})}{\eta_t} + \frac{L\eta_t}{2}\|g_t\|^2 + (\nabla f(x_t) - g_t)^\top \nabla f(x_t) \ .$$

Denoting $\Delta_t \coloneqq f(x_t) - f^*$ and $\Delta_{\max} \coloneqq \max_{t\in[T]} f(x_t) - f^*$. Thus, the above is equivalent to

$$\|\nabla f(x_t)\|^2 \leq \frac{\Delta_t - \Delta_{t+1}}{\eta_t} + \frac{L\eta_t}{2}\|g_t\|^2 + (\nabla f(x_t) - g_t)^\top \nabla f(x_t) \ .$$

Summing over $t = 1, \ldots, T$, we obtain that

$$\begin{aligned} \sum_{t\in[T]} \|\nabla f(x_t)\|^2 &\leq \sum_{t\in[T]} \frac{\Delta_t - \Delta_{t+1}}{\eta_t} + \frac{L}{2}\sum_{t\in[T]} \eta_t \|g_t\|^2 + \sum_{t\in[T]} (\nabla f(x_t) - g_t)^\top \nabla f(x_t) \\ &\leq \frac{\Delta_{\max}}{\eta_T} + \frac{L}{2}\sum_{t\in[T]} \eta_t \|g_t\|^2 + \sum_{t\in[T]} (\nabla f(x_t) - g_t)^\top \nabla f(x_t) \ , \end{aligned} \tag{30}$$

where the last inequality follows from:

$$\sum_{t=1}^{T} \frac{\Delta_t - \Delta_{t+1}}{\eta_t} = \frac{\Delta_1}{\eta_1} + \sum_{t=2}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \Delta_t - \frac{\Delta_{T+1}}{\eta_T} \leq \frac{\Delta_1}{\eta_1} + \sum_{t=2}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \Delta_t \leq \frac{\Delta_{\max}}{\eta_T} ,$$

which holds as $\eta_t = \eta_0 (\sum_{\tau=1}^{t} \|g_\tau\|^2)^{-1/2}$ is non-increasing. Next, we can bound the second term in the R.H.S of Equation (30) using Lemma I.4 with $a_i = \|g_i\|^2$ as follows:

$$\sum_{t \in [T]} \eta_t \|g_t\|^2 = \eta_0 \sum_{t \in [T]} \frac{\|g_t\|^2}{\sqrt{\sum_{s=1}^{t} \|g_s\|^2}} \leq 2\eta_0 \sqrt{\sum_{t \in [T]} \|g_t\|^2} .$$

Injecting this bound and η_T back into Equation (30) and considering that $\Delta_{\max} \leq 2M$ concludes the proof. □

Leveraging Lemma G.3, we derive the following bound, instrumental in proving Theorem 5.2.

Lemma G.4. *Assume f is bounded by M and consider* (AdaGrad-Norm) *with $\mathcal{K} = \mathbb{R}^d$. Denote:*

$$G_T^2 \coloneqq \sum_{t=1}^{T} \|\nabla f(x_t)\|^2, \quad V_{1:T}^2 \coloneqq \sum_{t=1}^{T} V_t^2, \quad S_T^2 \coloneqq \sum_{t=1}^{T} \|b_t\|^2 .$$

For every $T \geq 1$, it holds that

$$\mathbb{E} G_T^2 \leq 2\zeta \left(\sqrt{\mathbb{E} V_{1:T}^2} + \sqrt{2\mathbb{E} S_T^2} + \sqrt{2\mathbb{E} G_T^2} \right) + \mathbb{E} S_T^2 ,$$

where $\zeta \coloneqq \frac{2M}{\eta_0} + \eta_0 L$.

Proof. Employing Lemma G.3, we get

$$\mathbb{E} G_T^2 \leq \zeta \underbrace{\mathbb{E} \left[\sqrt{\sum_{t \in [T]} \|g_t\|^2} \right]}_{=(A)} + \underbrace{\sum_{t \in [T]} \mathbb{E}[-b_t^\top \nabla f(x_t)]}_{=(B)} .$$

Bounding (A). We apply a technique akin to the one utilized in the proof of Lemma G.2. Concretely, applying Equation (29) and using Jensen's inequality, yields

$$\mathbb{E} \left[\sqrt{\sum_{t \in [T]} \|g_t\|^2} \right] \leq \sqrt{\sum_{t \in [T]} \left(\mathbb{E} V_t^2 + 2\mathbb{E}\|b_t\|^2 + 2\mathbb{E}\|\nabla f(x_t)\|^2 \right)} \leq \sqrt{\mathbb{E} V_{1:T}^2} + \sqrt{2\mathbb{E} S_T^2} + \sqrt{2\mathbb{E} G_T^2} .$$

Bounding (B). Employing Young's inequality, namely, $a^\top b \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$, results in

$$\sum_{t \in [T]} \mathbb{E}[-b_t^\top \nabla f(x_t)] \leq \frac{1}{2} \sum_{t \in [T]} \mathbb{E}\|b_t\|^2 + \frac{1}{2} \sum_{t \in [T]} \mathbb{E}\|\nabla f(x_t)\|^2 = \frac{1}{2}\mathbb{E} S_T^2 + \frac{1}{2}\mathbb{E} G_T^2 .$$

Substituting the bounds on (A) and (B) implies that

$$\mathbb{E} G_T^2 \leq \zeta \left(\sqrt{\mathbb{E} V_{1:T}^2} + \sqrt{2\mathbb{E} S_T^2} + \sqrt{2\mathbb{E} G_T^2} \right) + \frac{1}{2}\mathbb{E} S_T^2 + \frac{1}{2}\mathbb{E} G_T^2 .$$

Subtracting $\frac{1}{2}\mathbb{E} G_T^2$ and multiplying by 2 establishes the result. □

H. Dynamic and Adaptive Byzantine-Robustness with MFM and AdaGrad

Following the methodology outlined in Appendix D, this section focuses on analyzing Algorithm 2 with **Option 2**. Here, we replace the general (δ, κ)-robust aggregator with the MFM aggregator and incorporate the adaptive AdaGrad learning rate (Levy et al., 2018; Ward et al., 2020; Attia & Koren, 2023).

Recall that Algorithm 2 with **Option 2** and learning rate defined in Equation (7) performs the following update rule for every $t \in [T]$:

$$J_t \sim \text{Geom}(1/2)$$

$$\widehat{g}_t^j \leftarrow \text{MFM}(\overline{g}_{t,1}^j, \dots, \overline{g}_{t,m}^j; \mathcal{T}^j), \quad \mathcal{T}^j := \frac{2C\mathcal{V}}{\sqrt{2^j}} = 4\mathcal{V}\sqrt{\frac{2\log(16m^2T)}{2^j}} \tag{31}$$

$$g_t \leftarrow \widehat{g}_t^0 + \begin{cases} 2_t^J \left(\widehat{g}_t^{J_t} - \widehat{g}_t^{J_t-1}\right), & \text{if } J_t \le J_{\max} := \lfloor \log T \rfloor \text{ and } \mathcal{E}_t(J_t) \text{ holds} \\ 0, & \text{otherwise} \end{cases} \tag{32}$$

$$x_{t+1} \leftarrow \Pi_{\mathcal{K}}(x_t - \eta_t g_t), \quad \eta_t = \frac{\eta_0}{\sqrt{\sum_{s=1}^t \|g_s\|^2}}, \tag{33}$$

where the associated event $\mathcal{E}_t(J_t)$ in this case is defined as,

$$\mathcal{E}_t(J_t) := \left\{ \|\widehat{g}_t^{J_t} - \widehat{g}_t^{J_t-1}\| \le (1+\sqrt{2})\frac{c_{\mathcal{E}} C\mathcal{V}}{\sqrt{2^{J_t}}} \right\}, \quad c_{\mathcal{E}} := 6\sqrt{2}, \quad \tilde{C} := 2\sqrt{2}C = 8\sqrt{\log(4m^2T)}\,. \tag{34}$$

For ease of writing, we denote: $\tilde{\gamma} := \frac{1}{m} + 32\delta^2$.

We start by showing that $\mathcal{E}_t$ is satisfied with high probability, mirroring Lemma D.3.

Lemma H.1. *Consider $\mathcal{E}(J_t)$ defined in Equation (34). For every $t \in \tau_s$ and $j = 0, \dots, J_{\max}$, we have*

$$\mathbb{P}_{t-1}(\mathcal{E}_t(j)) \ge 1 - \frac{1}{4mT},$$

where the randomness is w.r.t the stochastic gradient samples.

Proof. By item 2 of Lemma F.3, for every $j = 0, \dots, J_{\max}$ (separately), we have with probability at least $1 - \frac{1}{8mT}$,

$$\|\widehat{g}_t^j - \nabla_t\|^2 \le \frac{2C^2\mathcal{V}^2}{2^j}\left(\frac{1}{m} + 128\delta^2\right) \le \frac{2C^2\mathcal{V}^2}{2^j}\left(\frac{4}{m} + 128\delta^2\right) = \frac{8C^2\mathcal{V}^2\tilde{\gamma}}{2^j} = \frac{\tilde{C}^2\mathcal{V}^2\tilde{\gamma}}{2^j},$$

where we used $\frac{1}{m} + 128\delta^2 \le 4\left(\frac{1}{m} + 32\delta^2\right) = 4\tilde{\gamma}$ and $\tilde{C}^2 = 8C^2$. Therefore, we have

$$\mathbb{P}_{t-1}\left(\|\widehat{g}_t^j - \nabla_t\| > \frac{c_{\mathcal{E}} C\mathcal{V}}{\sqrt{2^j}}\right) = \mathbb{P}_{t-1}\left(\|\widehat{g}_t^j - \nabla_t\| > \frac{3\tilde{C}\mathcal{V}}{\sqrt{2^j}}\right) \le \mathbb{P}_{t-1}\left(\|\widehat{g}_t^j - \nabla_t\| > \frac{\tilde{C}\mathcal{V}\sqrt{\tilde{\gamma}}}{\sqrt{2^j}}\right) \le \frac{1}{8mT},$$

where we used $c_{\mathcal{E}}C = 6\sqrt{2}C = 3\tilde{C}$, and $\tilde{\gamma} = \frac{1}{m} + 32\delta^2 \le 9$ as $\delta < 1/2$. This bound, in conjunction with the union bound, allows us to bound $\mathbb{P}_{t-1}(\mathcal{E}_t(j)^c)$ as,

$$\begin{aligned}
\mathbb{P}_{t-1}(\mathcal{E}_t(j)^c) &= \mathbb{P}_{t-1}\left(\|\widehat{g}_t^j - \widehat{g}_t^{j-1}\| > (1+\sqrt{2})\frac{c_{\mathcal{E}} C\mathcal{V}}{\sqrt{2^j}}\right) \\
&\le \mathbb{P}_{t-1}\left(\left\{\|\widehat{g}_t^j - \nabla_t\| > \frac{c_{\mathcal{E}} C\mathcal{V}}{\sqrt{2^j}}\right\} \bigcup \left\{\|\widehat{g}_t^{j-1} - \nabla_t\| > \frac{c_{\mathcal{E}} C\mathcal{V}}{\sqrt{2^{j-1}}}\right\}\right) \\
&\le \mathbb{P}_{t-1}\left(\|\widehat{g}_t^j - \nabla_t\| > \frac{c_{\mathcal{E}} C\mathcal{V}}{\sqrt{2^j}}\right) + \mathbb{P}_{t-1}\left(\|\widehat{g}_t^{j-1} - \nabla_t\| > \frac{c_{\mathcal{E}} C\mathcal{V}}{\sqrt{2^{j-1}}}\right) \\
&\le \frac{1}{8mT} + \frac{1}{8mT} = \frac{1}{4mT}.
\end{aligned}$$

□

Next, using Corollary F.4 and Lemma H.1, we establish bounds on the bias and variance of the MLMC estimator defined in Equation (32), resembling those outlined in Lemma D.4.

Lemma H.2 (MLMC Bias and Variance)**.** *Consider g_t defined as in Equation (32). Then, for every $m \geq 2$,*

1. *The bias $b_t \coloneqq \mathbb{E}_{t-1} g_t - \nabla_t$ is bounded as*

$$\|b_t\| \leq \begin{cases} \tilde{C}\mathcal{V}\sqrt{\frac{2\tilde{\gamma}}{T}} + \frac{\sqrt{5}\tilde{C}\mathcal{V}\log T}{mT} + \frac{(1+\sqrt{2})\|\nabla_t\|}{2\sqrt{2mT}}, & t \in \tau_s \\ \sqrt{125\log T}\tilde{C}\mathcal{V} + \frac{\|\nabla_t\|}{2\sqrt{mT}}, & t \in \tau_d \end{cases} .$$

2. *The variance $V_t^2 \coloneqq \mathbb{E}_{t-1}\left\|g_t - \mathbb{E}_{t-1} g_t\right\|^2$ is bounded as*

$$V_t^2 \leq \begin{cases} 14\tilde{C}^2\mathcal{V}^2\tilde{\gamma}\log T + \log T\left\|\nabla_t\right\|^2, & t \in \tau_s \\ 125\tilde{C}^2\mathcal{V}^2\log T + \frac{\|\nabla_t\|^2}{4mT}, & t \in \tau_d \end{cases} .$$

Proof. Our proof technique parallels that of Lemma D.4. Starting in a similar fashion, we can bound the variance as shown in Equation (20), namely,

$$V_t^2 \leq \mathbb{E}_{t-1}\left\|g_t - \nabla_t\right\|^2 \leq 2\mathbb{E}_{t-1}\left\|\widehat{g}_t^0 - \nabla_t\right\|^2 + 2\sum_{j=1}^{J_{\max}} 2^j \underbrace{\mathbb{E}_{t-1}\left[\|\widehat{g}_t^j - \widehat{g}_t^{j-1}\|^2 \mathbb{1}_{\mathcal{E}_t(j)}\right]}_{=(\dagger)} . \tag{35}$$

Unlike in Lemma D.4, here we bound $(\dagger)$ differently for $t \in \tau_d$ and $t \in \tau_s$. This is because the bound within the event $\mathcal{E}_t$ in Equation (34) deviates from that in Equation (17) by a factor of $\sqrt{\gamma}$. Alternatively, one could introduce a factor of $\sqrt{\tilde{\gamma}}$ to maintain a similar analysis; however, in doing so, the event would no longer be oblivious to δ, which contradicts one of our objectives in utilizing the AdaGrad learning rate.

For every $t \in [T]$ (including $t \in \tau_d$), it holds that

$$\mathbb{E}_{t-1}\left[\|\widehat{g}_t^j - \widehat{g}_t^{j-1}\|^2 \mathbb{1}_{\mathcal{E}_t(j)}\right] = \mathbb{E}_{t-1}\left[\|\widehat{g}_t^j - \widehat{g}_t^{j-1}\|^2 | \mathcal{E}_t(j)\right] \underbrace{\mathbb{P}(\mathcal{E}_t(j))}_{\leq 1} \leq \frac{(1+\sqrt{2})^2 c_{\mathcal{E}}^2 C^2 \mathcal{V}^2}{2^j} \leq \frac{54\tilde{C}^2\mathcal{V}^2}{2^j} ,$$

where in the final inequality, we utilize the constraint on $\|\widehat{g}_t^j - \widehat{g}_t^{j-1}\|$, conditioned on the event $\mathcal{E}_t(j)$ (cf. Equation (34)). However, considering $t \in \tau_s$ (static rounds), we employ a more careful analysis. Specifically, Corollary F.4 implies that

$$\begin{aligned} \mathbb{E}_{t-1}\left[\|\widehat{g}_t^j - \widehat{g}_t^{j-1}\|^2 \mathbb{1}_{\mathcal{E}_t(j)}\right] \leq \mathbb{E}_{t-1}\|\widehat{g}_t^j - \widehat{g}_t^{j-1}\|^2 &\leq 2\mathbb{E}_{t-1}\|\widehat{g}_t^j - \nabla_t\|^2 + 2\mathbb{E}_{t-1}\|\widehat{g}_t^{j-1} - \nabla_t\|^2 \\ &\leq 2\left(\frac{\tilde{C}^2\mathcal{V}^2\tilde{\gamma}}{2^j} + \frac{\|\nabla_t\|^2}{8mT}\right) + 2\left(\frac{\tilde{C}^2\mathcal{V}^2\tilde{\gamma}}{2^{j-1}} + \frac{\|\nabla_t\|^2}{8mT}\right) \\ &= 2\left(\frac{3\tilde{C}^2\mathcal{V}^2\tilde{\gamma}}{2^j} + \frac{\|\nabla_t\|^2}{4mT}\right) \\ &\leq \frac{6\tilde{C}^2\mathcal{V}^2\tilde{\gamma} + \frac{1}{4}\|\nabla_t\|^2}{2^j} , \end{aligned}$$

where the last inequality follows from $2^j \leq T$ and $m \geq 2$. We can thus conclude that

$$\mathbb{E}_{t-1}\left[\|\widehat{g}_t^j - \widehat{g}_t^{j-1}\|^2 \mathbb{1}_{\mathcal{E}_t(j)}\right] \leq \frac{1}{2^j} \cdot \begin{cases} 54\tilde{C}^2\mathcal{V}^2, & t \in \tau_d \\ 6\tilde{C}^2\mathcal{V}^2\tilde{\gamma} + \frac{1}{4}\|\nabla_t\|^2, & t \in \tau_s \end{cases} .$$

In addition, Corollary F.4 implies that $\mathbb{E}_{t-1}\|\widehat{g}_t^0 - \nabla_t\|^2 \leq \tilde{C}^2\mathcal{V}^2\tilde{\gamma} + \frac{\|\nabla_t\|^2}{8mT}$. Plugging these bounds back into Equation (35) establishes the variance bound. Specifically, for $t \in \tau_d$,

$$\begin{aligned} V_t^2 \leq \mathbb{E}_{t-1}\left\|g_t - \nabla_t\right\|^2 &\leq 2\left(\tilde{C}^2\mathcal{V}^2\tilde{\gamma} + \frac{\|\nabla_t\|^2}{8mT}\right) + 2\sum_{j=1}^{J_{\max}} 2^j \cdot \frac{54\tilde{C}^2\mathcal{V}^2}{2^j} \\ &= 2\tilde{C}^2\mathcal{V}^2\tilde{\gamma} + \frac{\|\nabla_t\|^2}{4mT} + 108\tilde{C}^2\mathcal{V}^2 J_{\max} \\ &\leq 125\tilde{C}^2\mathcal{V}^2\log T + \frac{\|\nabla_t\|^2}{4mT}\ , \end{aligned}$$

where in the last inequality we used $\tilde{\gamma} \leq 8.5$ for $m \geq 2$, and $1 \leq J_{\max} \leq \log T$. On the other hand, for $t \in \tau_s$, it holds that

$$\begin{aligned} V_t^2 \leq \mathbb{E}_{t-1}\left\|g_t - \nabla_t\right\|^2 &\leq 2\left(\tilde{C}^2\mathcal{V}^2\tilde{\gamma} + \frac{\|\nabla_t\|^2}{8mT}\right) + 2\sum_{j=1}^{J_{\max}} 2^j \cdot \frac{6\tilde{C}^2\mathcal{V}^2\tilde{\gamma} + \frac{1}{4}\|\nabla_t\|^2}{2^j} \\ &= 2\tilde{C}^2\mathcal{V}^2\tilde{\gamma} + \frac{\|\nabla_t\|^2}{4mT} + \left(12\tilde{C}^2\mathcal{V}^2\tilde{\gamma} + \frac{1}{2}\|\nabla_t\|^2\right) J_{\max} \\ &\leq 14\tilde{C}^2\mathcal{V}^2\tilde{\gamma}\log T + \log T\,\|\nabla_t\|^2\ , \end{aligned}$$

where the last inequality follows from $1 \leq J_{\max} \leq \log T$ and $\frac{1}{2mT} \leq \log T$.

Moving forward, we proceed to establish a bounds on the squared bias, following a similar approach as demonstrated in the proof of Lemma D.4. For $t \in \tau_d$, we trivially bound the bias by the square root of the MSE as follows:

$$\|b_t\| = \|\mathbb{E}_{t-1}g_t - \nabla_t\| \leq \sqrt{\mathbb{E}_{t-1}\|g_t - \nabla_t\|^2} \leq \sqrt{125\tilde{C}^2\mathcal{V}^2\log T + \frac{\|\nabla_t\|^2}{4mT}} \leq \sqrt{125\log T}\tilde{C}\mathcal{V} + \frac{\|\nabla_t\|}{2\sqrt{mT}}\ . \tag{36}$$

For $t \in \tau_s$, we repeat the steps from the proof of Lemma D.4, leading to the derivation of Equation (22), i.e.,

$$\|b_t\| \leq \|\mathbb{E}_{t-1}[\widehat{g}_t^{J_{\max}} - \nabla_t]\| + \|\mathbb{E}_{t-1}y_t\| \leq \underbrace{\|\mathbb{E}_{t-1}[\widehat{g}_t^{J_{\max}} - \nabla_t]\|}_{=(A)} + \underbrace{\mathbb{E}_{t-1}\|y_t\|}_{=(B)}\ , \tag{37}$$

where $y_t = -\sum_{j=1}^{J_{\max}} z_t^j$, and $z_t^j = \mathbb{E}_{t-1}\left[\widehat{g}_t^j - \widehat{g}_t^{j-1}|\mathcal{E}_t(j)^c\right]\mathbb{P}_{t-1}(\mathcal{E}_t(j)^c)$.

Bounding (A): Utilizing Corollary F.4 and Jensen's inequality, we get

$$\|\mathbb{E}_{t-1}[\widehat{g}_t^{J_{\max}} - \nabla_t]\| \leq \sqrt{\mathbb{E}_{t-1}\|\widehat{g}_t^{J_{\max}} - \nabla_t\|^2} \leq \sqrt{\frac{\tilde{C}^2\mathcal{V}^2\tilde{\gamma}}{2^{J_{\max}}} + \frac{\|\nabla_t\|^2}{8mT}} \leq \tilde{C}\mathcal{V}\sqrt{\frac{2\tilde{\gamma}}{T}} + \frac{\|\nabla_t\|}{\sqrt{8mT}}\ ,$$

where we used $2^{J_{\max}} \geq T/2$.

Bounding (B): By the triangle inequality and Jensen's inequality, we have

$$\|y_t\| \leq \sum_{j=1}^{J_{\max}} \|z_t^j\| \leq \sum_{j=1}^{J_{\max}} \mathbb{E}_{t-1}\Big[\|\widehat{g}_t^j - \widehat{g}_t^{j-1}\||\mathcal{E}_t(j)^c\Big]\mathbb{P}_{t-1}(\mathcal{E}_t(j)^c)\ .$$

By item 1 of Lemma F.3, for every $j = 0, \ldots, J_{\max}$, we have that $\|\widehat{g}_t^j - \nabla_t\| \leq \sqrt{20C^2\mathcal{V}^2 + \|\nabla_t\|^2} \leq \sqrt{20}C\mathcal{V} + \|\nabla_t\|$. Therefore, it holds that

$$\|\widehat{g}_t^j - \widehat{g}_t^{j-1}\| \leq \|\widehat{g}_t^j - \nabla_t\| + \|\widehat{g}_t^{j-1} - \nabla_t\| \leq 2\left(\sqrt{20}C\mathcal{V} + \|\nabla_t\|\right)\ .$$

In addition, by Lemma H.1, we have for every $j = 0, \dots, J_{\max}$ that $\mathbb{P}_{t-1}(\mathcal{E}_t(j)^c) \leq \frac{1}{4mT}$. Combining these bounds, we conclude that

$$\|y_t\| \leq \sum_{j=1}^{J_{\max}} 2\left(\sqrt{20}C\mathcal{V} + \|\nabla_t\|\right) \cdot \frac{1}{4mT} = \frac{(\sqrt{20}C\mathcal{V} + \|\nabla_t\|)J_{\max}}{2mT} \leq \frac{(\sqrt{20}C\mathcal{V} + \|\nabla_t\|)\log T}{2mT} .$$

Substituting the bounds on (A) and (B) back into Equation (37) implies that for every $t \in \tau_s$:

$$\begin{aligned}
\|b_t\| \leq \tilde{C}\mathcal{V}\sqrt{\frac{2\tilde{\gamma}}{T}} + \frac{\|\nabla_t\|}{\sqrt{8mT}} + \frac{(\sqrt{20}\tilde{C}\mathcal{V} + \|\nabla_t\|)\log T}{2mT} &= \tilde{C}\mathcal{V}\sqrt{\frac{2\tilde{\gamma}}{T}} + \frac{\sqrt{5}\tilde{C}\mathcal{V}\log T}{mT} + \frac{\|\nabla_t\|}{2}\left(\frac{1}{\sqrt{2mT}} + \frac{\log T}{mT}\right) \\
&\leq \tilde{C}\mathcal{V}\sqrt{\frac{2\tilde{\gamma}}{T}} + \frac{\sqrt{5}\tilde{C}\mathcal{V}\log T}{mT} + \frac{\|\nabla_t\|}{2\sqrt{2m}}\left(\frac{1}{\sqrt{T}} + \frac{\log T}{T}\right) \\
&\leq \tilde{C}\mathcal{V}\sqrt{\frac{2\tilde{\gamma}}{T}} + \frac{\sqrt{5}\tilde{C}\mathcal{V}\log T}{mT} + \frac{(1+\sqrt{2})\,\|\nabla_t\|}{2\sqrt{2mT}} ,
\end{aligned}$$

where the second inequality follows from $\frac{1}{m} \leq \frac{1}{\sqrt{2m}}$, which holds for every $m \geq 2$; and in the last inequality we used $\log T \leq \sqrt{2T}$, which holds $\forall T \geq 1$. □

Similarly to the approach employed in Appendix D, we now utilize the established bias and variance bounds to derive convergence guarantees for Algorithm 2 with **Option 2**. We use the following notations, as in Appendix G:

$$\mathcal{R}_T := \sum_{t\in[T]} (f(x_t) - f(x^*)),\ \ G_T^2 := \sum_{t\in[T]} \|\nabla f(x_t)\|^2,\ \ V_{1:T}^2 := \sum_{t\in[T]} V_t^2,\ \ b_{1:T} := \sum_{t\in[T]} \|b_t\|,\ \ S_T^2 := \sum_{t\in[T]} \|b_t\|^2 .$$

H.1. Convex Case

The following theorem implies convergence in the convex case. For ease of analysis, we assume that $\nabla f(x^*) = 0$, which enables using Lemma I.3; this is the case when $\mathcal{K}$ contains the global minimizer of f. To alleviate this assumption, one could consider adopting a more sophisticated *optimistic* approach (Mohri & Yang, 2016). Yet, we refrain from doing so to maintain the clarity of our presentation and analysis.

Theorem H.3. *Assume f is convex and x^* satisfies $\nabla f(x^*) = 0$. Under Assumptions 2.2 and 2.3, consider Algorithm 2 with **Option 2** and the AdaGrad-Norm learning rate specified in Equation (7), where $\eta_0 = D/2$. For every $T \geq 1$, we have*

$$\mathbb{E}f(\bar{x}_T) - f(x^*) \leq \tilde{C}D\mathcal{V}\left(14\sqrt{\frac{\tilde{\gamma}\log T}{T}} + \frac{24|\tau_d|\sqrt{\log T}}{T} + \frac{78\sqrt{|\tau_d|\log T}}{T} + \frac{6\log T}{mT} + \frac{10\sqrt{\tilde{\gamma}}}{T} + \frac{16\log T}{mT^{3/2}}\right) + \frac{392LD^2\log T}{T} ,$$

where $\bar{x}_T := \frac{1}{T}\sum_{t=1}^T x_t$ and $\tilde{C} := 2\sqrt{2}C$.

Proof. Applying Lemma G.2, we have

$$\begin{aligned}
\mathbb{E}\mathcal{R}_T &\leq D\sqrt{2\sum_{t\in[T]} \mathbb{E}V_t^2} + 2D\sqrt{\sum_{t\in[T]} \mathbb{E}\|b_t\|^2} + D\sum_{t\in[T]} \mathbb{E}\|b_t\| + 2D\sqrt{\sum_{t\in[T]} \mathbb{E}\|\nabla f(x_t)\|^2} \\
&= D\sqrt{2\mathbb{E}V_{1:T}^2} + D\mathbb{E}b_{1:T} + 2D\sqrt{\mathbb{E}S_T^2} + 2D\sqrt{\mathbb{E}G_T^2} . \qquad (38)
\end{aligned}$$

Bounding $\mathbb{E}V_{1:T}^2$. Utilizing the variance bound from Lemma H.2, we get that

$$\begin{aligned}
\mathbb{E}V_{1:T}^2 &= \sum_{t\in\tau_d} \mathbb{E}V_t^2 + \sum_{t\in\tau_s} \mathbb{E}V_t^2 \\
&\leq \sum_{t\in\tau_d}\left(125\tilde{C}^2\mathcal{V}^2\log T + \frac{\mathbb{E}\|\nabla f(x_t)\|^2}{4mT}\right) + \sum_{t\in\tau_s}\left(14\tilde{C}^2\mathcal{V}^2\tilde{\gamma}\log T + \log T\mathbb{E}\|\nabla f(x_t)\|^2\right) \\
&\leq 125\tilde{C}^2\mathcal{V}^2|\tau_d|\log T + 14\tilde{C}^2\mathcal{V}^2\tilde{\gamma}T\log T + \log T\mathbb{E}G_T^2 . \qquad (39)
\end{aligned}$$

Bounding $\mathbb{E}b_{1:T}$. Employing the bias bound from Lemma H.2, we obtain:

$$\begin{aligned}
\mathbb{E}b_{1:T} &= \sum_{t\in\tau_d}\mathbb{E}\|b_t\| + \sum_{t\in\tau_s}\mathbb{E}\|b_t\| \\
&\le \sum_{t\in\tau_d}\left(\sqrt{125\log T}\tilde{C}\mathcal{V} + \frac{\mathbb{E}\|\nabla f(x_t)\|}{2\sqrt{mT}}\right) + \sum_{t\in\tau_s}\left(\tilde{C}\mathcal{V}\sqrt{\frac{2\tilde{\gamma}}{T}} + \frac{\sqrt{5}\tilde{C}\mathcal{V}\log T}{mT} + \frac{(1+\sqrt{2})\mathbb{E}\|\nabla f(x_t)\|}{2\sqrt{2mT}}\right) \\
&\le \sqrt{125\log T}\tilde{C}\mathcal{V}|\tau_d| + \tilde{C}\mathcal{V}\sqrt{2\tilde{\gamma}T} + \frac{\sqrt{5}\tilde{C}\mathcal{V}\log T}{m} + \frac{1+\sqrt{2}}{2\sqrt{2mT}}\sum_{t\in[T]}\mathbb{E}\|\nabla f(x_t)\| \\
&\le \sqrt{125\log T}\tilde{C}\mathcal{V}|\tau_d| + \tilde{C}\mathcal{V}\sqrt{2\tilde{\gamma}T} + \frac{\sqrt{5}\tilde{C}\mathcal{V}\log T}{m} + \frac{1+\sqrt{2}}{2\sqrt{2m}}\sqrt{\mathbb{E}G_T^2}\ ,
\end{aligned}$$

where in the second inequality we used the fact that $\frac{1+\sqrt{2}}{\sqrt{2}} \ge 1$, and the last inequality arises from the application of the Cauchy-Schwarz inequality and Jensen's inequality, specifically $\sum_{t\in[T]}\mathbb{E}\|\nabla f(x_t)\| \le \sqrt{T}\sqrt{\sum_{t\in[T]}\mathbb{E}\|\nabla f(x_t)\|^2}$.

Bounding $\mathbb{E}S_T^2$. We start by bounding $\|b_t\|^2$ for every $t\in[T]$. From Equation (36), we have for all $t\in\tau_d$ that

$$\|b_t\|^2 \le 125\tilde{C}^2\mathcal{V}^2\log T + \frac{\|\nabla f(x_t)\|^2}{4mT}\ .$$

For $t\in\tau_s$, employing Lemma H.2 and using $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ gives:

$$\|b_t\|^2 \le \left(\tilde{C}\mathcal{V}\sqrt{\frac{2\tilde{\gamma}}{T}} + \frac{\sqrt{5}\tilde{C}\mathcal{V}\log T}{mT} + \frac{(1+\sqrt{2})\,\|\nabla f(x_t)\|}{2\sqrt{2mT}}\right)^2 \le \frac{6\tilde{C}^2\mathcal{V}^2\tilde{\gamma}}{T} + \frac{15\tilde{C}^2\mathcal{V}^2\log^2 T}{m^2T^2} + \frac{9\mathbb{E}\|\nabla f(x_t)\|^2}{4mT}\ .$$

Thus, we can bound $\mathbb{E}S_T^2$ as follows:

$$\begin{aligned}
\mathbb{E}S_T^2 &= \sum_{t\in\tau_d}\mathbb{E}\|b_t\|^2 + \sum_{t\in\tau_s}\mathbb{E}\|b_t\|^2 \\
&\le \sum_{t\in\tau_d}\left(125\tilde{C}^2\mathcal{V}^2\log T + \frac{\mathbb{E}\|\nabla f(x_t)\|^2}{4mT}\right) + \sum_{t\in\tau_s}\left(\frac{6\tilde{C}^2\mathcal{V}^2\tilde{\gamma}}{T} + \frac{15\tilde{C}^2\mathcal{V}^2\log^2 T}{m^2T^2} + \frac{9\mathbb{E}\|\nabla f(x_t)\|^2}{4mT}\right) \\
&\le 125\tilde{C}^2\mathcal{V}^2|\tau_d|\log T + 6\tilde{C}^2\mathcal{V}^2\tilde{\gamma} + \frac{15\tilde{C}^2\mathcal{V}^2\log^2 T}{m^2T} + \frac{9}{4mT}\mathbb{E}G_T^2\ . \qquad (40)
\end{aligned}$$

Plugging these bound back into Equation (38) and rearranging terms yields:

$$\begin{aligned}
\mathbb{E}\mathcal{R}_T \le\ & D\sqrt{250\tilde{C}^2\mathcal{V}^2|\tau_d|\log T + 28\tilde{C}^2\mathcal{V}^2\tilde{\gamma}T\log T + 2\log T\mathbb{E}G_T^2} + \\
& \sqrt{125\log T}\tilde{C}D\mathcal{V}|\tau_d| + \tilde{C}D\mathcal{V}\sqrt{2\tilde{\gamma}T} + \frac{\sqrt{5}\tilde{C}D\mathcal{V}\log T}{m} + \frac{(1+\sqrt{2})D}{2\sqrt{2m}}\sqrt{\mathbb{E}G_T^2} + \\
& 2D\sqrt{125\tilde{C}^2\mathcal{V}^2|\tau_d|\log T + 6\tilde{C}^2\mathcal{V}^2\tilde{\gamma} + \frac{15\tilde{C}^2\mathcal{V}^2\log^2 T}{m^2T} + \frac{9}{4mT}\mathbb{E}G_T^2} + \\
& 2D\sqrt{\mathbb{E}G_T^2} \\
\le\ & (\sqrt{250}+2\sqrt{125})\tilde{C}D\mathcal{V}\sqrt{|\tau_d|\log T} + (\sqrt{28}+\sqrt{2})\tilde{C}D\mathcal{V}\sqrt{\tilde{\gamma}T\log T} + \sqrt{125}\tilde{C}D\mathcal{V}|\tau_d|\sqrt{\log T} + \\
& \sqrt{5}\tilde{C}D\mathcal{V}\frac{\log T}{m} + 2\sqrt{6}\tilde{C}D\mathcal{V}\sqrt{\tilde{\gamma}} + 2\sqrt{15}\tilde{C}D\mathcal{V}\frac{\log T}{m\sqrt{T}} + D\left(\sqrt{2\log T} + \frac{1+\sqrt{2}}{\sqrt{2m}} + \frac{3}{\sqrt{mT}} + 2\right)\sqrt{\mathbb{E}G_T^2} \\
\le\ & \underbrace{39\tilde{C}D\mathcal{V}\sqrt{|\tau_d|\log T} + 7\tilde{C}D\mathcal{V}\sqrt{\tilde{\gamma}T\log T} + 12\tilde{C}D\mathcal{V}|\tau_d|\sqrt{\log T} + 3\tilde{C}D\mathcal{V}\frac{\log T}{m} + 5\tilde{C}D\mathcal{V}\sqrt{\tilde{\gamma}} + 8\tilde{C}D\mathcal{V}\frac{\log T}{m\sqrt{T}}}_{:=B} + \\
& 7D\sqrt{2L\log T\mathbb{E}\mathcal{R}_T}\ ,
\end{aligned}$$

where in the last inequality follows from $\sqrt{2} + \frac{1+\sqrt{2}}{\sqrt{2m}} + \frac{3}{\sqrt{mT}} + 2 \leq 7$ (for $m \geq 2$), and the application of Lemma I.3, which holds since we assume $\mathcal{K}$ contains a global minimum, i.e., $\nabla f(x^*) = 0$. Employing Lemma I.8 with $a = \mathbb{E}\mathcal{R}_T$, $b = B$, $c = 7D$, and $d = 2L \log T$, we get

$$\mathbb{E}\mathcal{R}_T \leq 2B + 392LD^2 \log T \ .$$

Finally, dividing by T and utilizing Jensen's inequality establishes the result:

$$\mathbb{E}f(\bar{x}_T) - f(x^*) \leq \frac{\mathbb{E}\mathcal{R}_T}{T} \leq \tilde{C}D\mathcal{V}\left(14\sqrt{\frac{\tilde{\gamma}\log T}{T}} + \frac{24|\tau_d|\sqrt{\log T}}{T} + \frac{78\sqrt{|\tau_d|\log T}}{T} + \frac{6\log T}{mT} + \frac{10\sqrt{\tilde{\gamma}}}{T} + \frac{16\log T}{mT^{3/2}}\right) + \frac{392LD^2\log T}{T} \ .$$

□

Theorem H.3 suggests the following observation holds true.

Corollary H.4. *As long as $|\tau_d| \in \mathcal{O}(\sqrt{\tilde{\gamma}T})$, the first term dominates the convergence rate, implying an asymptotically optimal bound of $\tilde{\mathcal{O}}\big(D\mathcal{V}\sqrt{(\delta^2 + 1/m)/T}\big)$.*

H.2. Non-convex Case

Theorem 5.2. *Suppose Assumption 2.2 holds, with f bounded by M (i.e., $\max_x |f(x)| \leq M$). Define $\zeta := \frac{2M}{\eta_0} + \eta_0 L$, and consider Algorithm 2 with **Option 2** and the AdaGrad-Norm learning rate. For every $T \geq 3$,*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(x_t)\|^2 \leq 8\zeta\tilde{C}\mathcal{V}\left(\frac{27\sqrt{|\tau_d|\log T}}{T} + 8\sqrt{\frac{\tilde{\gamma}\log T}{T}} + \frac{6\log T}{mT^{3/2}}\right) + 4\tilde{C}^2\mathcal{V}^2\left(\frac{125|\tau_d|\log T}{T} + \frac{6\tilde{\gamma}}{T} + \frac{15\log^2 T}{m^2T^2}\right) + \frac{1024\zeta^2\log T}{T} \ .$$

Proof. Utilizing Lemma G.4 gives:

$$\mathbb{E}G_T^2 \leq 2\zeta\left(\sqrt{\mathbb{E}V_{1:T}^2} + \sqrt{2\mathbb{E}S_T^2} + \sqrt{2\mathbb{E}G_T^2}\right) + \mathbb{E}S_T^2 \ . \tag{41}$$

Employing the bounds on $\mathbb{E}V_{1:T}^2$ and $\mathbb{E}S_T^2$ as given in Equations (39) and (40), respectively, we obtain

$$\begin{aligned}
\mathbb{E}G_T^2 &\leq 2\zeta\Bigg(\sqrt{125\tilde{C}^2\mathcal{V}^2|\tau_d|\log T + 14\tilde{C}^2\mathcal{V}^2\tilde{\gamma}T\log T + \log T\mathbb{E}G_T^2} + \\
&\quad \sqrt{250\tilde{C}^2\mathcal{V}^2|\tau_d|\log T + 12\tilde{C}^2\mathcal{V}^2\tilde{\gamma} + \frac{30\tilde{C}^2\mathcal{V}^2\log^2 T}{m^2T} + \frac{9}{2mT}\mathbb{E}G_T^2} + \sqrt{2\mathbb{E}G_T^2}\Bigg) + \\
&\quad 125\tilde{C}^2\mathcal{V}^2|\tau_d|\log T + 6\tilde{C}^2\mathcal{V}^2\tilde{\gamma} + \frac{15\tilde{C}^2\mathcal{V}^2\log^2 T}{m^2T} + \frac{9}{4mT}\mathbb{E}G_T^2 \\
&\leq 2\zeta\tilde{C}\mathcal{V}\left(\left(\sqrt{125} + \sqrt{250}\right)\sqrt{|\tau_d|\log T} + 4\sqrt{\tilde{\gamma}T\log T} + 4\sqrt{\tilde{\gamma}} + \frac{6\log T}{m\sqrt{T}}\right) + \\
&\quad \tilde{C}^2\mathcal{V}^2\left(125|\tau_d|\log T + 6\tilde{\gamma} + \frac{15\log^2 T}{m^2T}\right) + \\
&\quad 2\zeta\left(\sqrt{\log T} + \frac{3}{\sqrt{2mT}} + \sqrt{2}\right)\sqrt{\mathbb{E}G_T^2} + \frac{9}{4mT}\mathbb{E}G_T^2 \\
&\leq 2\zeta\tilde{C}\mathcal{V}\left(27\sqrt{|\tau_d|\log T} + 8\sqrt{\tilde{\gamma}T\log T} + \frac{6\log T}{m\sqrt{T}}\right) + \tilde{C}^2\mathcal{V}^2\left(125|\tau_d|\log T + 6\tilde{\gamma} + \frac{15\log^2 T}{m^2T}\right) + \\
&\quad 8\zeta\sqrt{\log T\mathbb{E}G_T^2} + \frac{1}{2}\mathbb{E}G_T^2 \ ,
\end{aligned}$$

where in the last inequality we used $1 + \frac{3}{\sqrt{2mT}} + \sqrt{2} \leq 4$ and $\frac{9}{4mT} \leq \frac{1}{2}$, both of which hold for $mT \geq 6$. Subtracting $\frac{1}{2}\mathbb{E}G_T^2$ and multiplying by 2 gives:

$$\mathbb{E}G_T^2 \leq \underbrace{4\zeta\tilde{C}\mathcal{V}\left(27\sqrt{|\tau_d|\log T} + 8\sqrt{\tilde{\gamma}T\log T} + \frac{6\log T}{m\sqrt{T}}\right) + 2\tilde{C}^2\mathcal{V}^2\left(125|\tau_d|\log T + 6\tilde{\gamma} + \frac{15\log^2 T}{m^2 T}\right)}_{:=B} + 16\zeta\sqrt{\log T\mathbb{E}G_T^2} \ .$$

Similarly to the proof of Lemma G.2, we apply Lemma I.8 with $a = \mathbb{E}G_T^2$, $b = B$, $c = 16\zeta$, and $d = \log T$ to obtain:

$$\mathbb{E}G_T^2 \leq 2B + 1024\zeta^2 \log T \ .$$

Dividing by T concludes the proof,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(x_t)\|^2 = \frac{\mathbb{E}G_T^2}{T} \leq 8\zeta\tilde{C}\mathcal{V}\left(\frac{27\sqrt{|\tau_d|\log T}}{T} + 8\sqrt{\frac{\tilde{\gamma}\log T}{T}} + \frac{6\log T}{mT^{3/2}}\right) + 4\tilde{C}^2\mathcal{V}^2\left(\frac{125|\tau_d|\log T}{T} + \frac{6\tilde{\gamma}}{T} + \frac{15\log^2 T}{m^2T^2}\right) + \frac{1024\zeta^2\log T}{T} \ .$$

□

The above convergence bounds implies the subsequent result, mirroring Corollary 4.2.

Corollary H.5. *Theorem 5.2 establishes the following asymptotic convergence rate:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\|\nabla f(x_t)\right\|^2 \in \tilde{\mathcal{O}}\left(\zeta\mathcal{V}\sqrt{\frac{1}{T}\left(\delta^2 + \frac{1}{m}\right)} + \mathcal{V}^2\frac{|\tau_d|}{T}\right) \ .$$

Thus, as long as the number of bad rounds $|\tau_d|$ is $\tilde{\mathcal{O}}(\sqrt{(\delta^2 + 1/m)T})$ (omitting the dependence on η_0, L, M, and $\mathcal{V}$), the established convergence rate is asymptotically optimal.

I. Technical Lemmata

In this section, we provide all technical results required for our analysis.

The following result by Pinelis (1994) is a concentration inequality for bounded martingale difference sequence.

Lemma I.1 (Alistarh et al., 2018, Lemma 2.4)**.** *Let $X_1, \ldots, X_T \in \mathbb{R}^d$ be a random process satisfying $\mathbb{E}[X_t|X_1, \ldots, X_{t-1}] = 0$ and $\|X_t\| \leq M$ a.s. for all $t \in [T]$. Then, with probability at least $1 - p$:*

$$\left\|\sum_{t=1}^{T} X_t\right\| \leq M\sqrt{2T\log(2/p)} \ .$$

In our convex analysis, we use the following classical result for projected SGD.

Lemma I.2 (Alistarh et al., 2018, Fact 2.5)**.** *If $x_{t+1} = \Pi_{\mathcal{K}}(x_t - \eta g_t) := \arg\min_{y\in\mathcal{K}}\|y - (x_t - \eta g_t)\|^2$, then for every $x \in \mathcal{K}$, we have*

$$g_t^\top(x_t - x) \leq g_t^\top(x_t - x_{t+1}) - \frac{\|x_t - x_{t+1}\|^2}{2\eta} + \frac{\|x_t - x\|^2}{2\eta} - \frac{\|x_{t+1} - x\|^2}{2\eta} \ .$$

Next is a classical result for smooth functions.

Lemma I.3 (Levy et al., 2018, Lemma 4.1; Attia & Koren, 2023, Lemma 8)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an L-smooth function and $x^* = \arg\min_{x\in\mathbb{R}^d} f(x)$. Then,*

$$\|\nabla f(x)\|^2 \leq 2L\left(f(x) - f(x^*)\right), \quad \forall x \in \mathbb{R}^d \ .$$

We utilize the following lemma by Auer et al. (2002), commonly used in the online learning literature, when analyzing our method with the AdaGrad-Norm learning rate.

Lemma I.4 (McMahan & Streeter, 2010, Lemma 5; Levy, 2017, Lemma A.1)**.** *For any sequence $a_1, \ldots, a_n \in \mathbb{R}_+$,*

$$\sum_{i=1}^{n} \frac{a_i}{\sqrt{\sum_{j=1}^{i} a_j}} \leq 2\sqrt{\sum_{i=1}^{n} a_i} \ .$$

The next two lemmas arise from fundamental probability calculations.

Lemma I.5. *For any random variable X and event $\mathcal{E}$,*

$$\mathbb{E}[X \cdot \mathbb{1}_{\mathcal{E}}] = \mathbb{E}[X] - \mathbb{E}[X|\mathcal{E}^c]\mathbb{P}(\mathcal{E}^c) \ .$$

Proof. By the law of total expectation:

$$\mathbb{E}[X \cdot \mathbb{1}_{\mathcal{E}}] = \mathbb{E}[X \cdot \mathbb{1}_{\mathcal{E}}|\mathcal{E}]\mathbb{P}(\mathcal{E}) + \underbrace{\mathbb{E}[X \cdot \mathbb{1}_{\mathcal{E}}|\mathcal{E}^c]}_{=0}\mathbb{P}(\mathcal{E}^c) = \mathbb{E}[X|\mathcal{E}]\mathbb{P}(\mathcal{E}) = \mathbb{E}[X] - \mathbb{E}[X|\mathcal{E}^c]\mathbb{P}(\mathcal{E}^c) \ .$$

□

Lemma I.6. *For any three events A, B, C satisfying $\mathbb{P}(A) > 0$ and $\mathbb{P}(B|A) \geq \mathbb{P}(C|A)$, we have $\mathbb{P}(B) \geq \mathbb{P}(C) - \mathbb{P}(A^c)$.*

Proof. By the law of total probability, we have

$$\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) + \underbrace{\mathbb{P}(B|A^c)\mathbb{P}(A^c)}_{\geq 0} \geq \mathbb{P}(C|A)\mathbb{P}(A) \ . \tag{42}$$

Again, by the law of total probability,

$$\mathbb{P}(C) = \mathbb{P}(C|A)\mathbb{P}(A) + \underbrace{\mathbb{P}(C|A^c)}_{\leq 1}\mathbb{P}(A^c) \leq \mathbb{P}(C|A)\mathbb{P}(A) + \mathbb{P}(A^c) \ .$$

Since $\mathbb{P}(A) > 0$, we can establish a lower bound for $\mathbb{P}(C|A)$ as follows:

$$\mathbb{P}(C|A) \geq \frac{\mathbb{P}(C) - \mathbb{P}(A^c)}{\mathbb{P}(A)} \ .$$

Substituting this bound back gives:

$$\mathbb{P}(B) \geq \frac{\mathbb{P}(C) - \mathbb{P}(A^c)}{\mathbb{P}(A)} \cdot \mathbb{P}(A) = \mathbb{P}(C) - \mathbb{P}(A^c) \ .$$

□

Finally, we utilize the following lemmas in our analysis to establish convergence rates.

Lemma I.7. *Let $a, b \geq 0$, $c > 0$, and consider $\eta = \min\left\{\sqrt{a/b}, 1/c\right\}$. Then,*

$$\frac{a}{\eta} + b\eta \leq 2\sqrt{ab} + ac \ .$$

Proof. Assume that $\sqrt{a/b} \leq 1/c$. In this case we have $\frac{a}{\eta} + b\eta = 2\sqrt{ab}$. Alternatively, if $1/c \leq \sqrt{a/b}$, then $\frac{a}{\eta} + b\eta = ac + b/c \leq ac + \sqrt{ab}$. Therefore, we always have

$$\frac{a}{\eta} + b\eta \leq \max\left\{2\sqrt{ab}, ac + \sqrt{ab}\right\} \leq 2\sqrt{ab} + ac \ .$$

□

Lemma I.8. *Let $a, b, c, d \geq 0$ with $a > 0$. If $a \leq b + c\sqrt{da}$, then $a \leq 2b + 4dc^2$*

Proof. Consider two cases. If $b \geq c\sqrt{d \cdot a}$, then we can bound,

$$a \leq b + c\sqrt{d \cdot a} \leq 2b \ .$$

Otherwise, $b < c\sqrt{d \cdot a}$ and we can bound,

$$a \leq b + c\sqrt{d \cdot a} \leq 2c\sqrt{d \cdot a} \ .$$

Dividing by $\sqrt{a} > 0$, we get that $\sqrt{a} \leq 2c\sqrt{d}$, which is equivalent to $a \leq 4dc^2$. We can thus conclude that

$$a \leq \max\left\{2b, 4dc^2\right\} \leq 2b + 4dc^2 \ .$$

□

J. Experimental Details

In this section, we describe the experimental setup and training details for the image classification tasks in Section 6.

Hardware and training times. We run all experiments on a machine with a single NVIDIA GeForce RTX 4090 GPU. For the MNIST experiments, the runtime of each configuration is ≈ 5 minutes and for CIFAR-10 it is ≈ 45 minutes.

Architectures and training details. We adopt the CNN architectures from Allouah et al. (2023), as detailed in Table 2.

Table 2. Training details and hyperparameters.

Dataset	MNIST	CIFAR-10
Architecture	Conv(20)-ReLU-MaxPool-Conv(20)-ReLU-MaxPool-FC(500)-ReLU-FC(10)-SoftMax	Conv(64)-ReLU-BatchNorm-Conv(64)-ReLU-BatchNorm-MaxPool-Dropout(0.25)-Conv(128)-ReLU-BatchNorm-Conv(128)-ReLU-BatchNorm-MaxPool-Dropout(0.25)-FC(128)-ReLU-Dropout(0.25)-FC(10)-Softmax
# of iterations	5000	8000
Learning rate	$\times 10$ drop after 4000 iterations	$\times 10$ drop after 6000 iterations
Weight decay	10^{-4}	
Base batch size	32	64
# of workers (m)	17	25

In our experiments, we use a base mini-batch of size B (32 for the MNIST experiments and 64 for the CIFAR-10 experiments). That is, in each iteration, each worker observes B samples. This implies that for the baselines we use a fixed mini-batch of size B, whereas for our method, which employs the MLMC estimator, in level J we use $B \cdot 2^J$ samples for gradient estimation. Following Dorfman & Levy (2022), we limit the maximal value of J to be smaller than $\lfloor \log T \rfloor$, specifically $J_{\max} = 7$, which we found to perform well in practice. For both MNIST and CIFAR-10 we use a learning rate schedule where the initial learning rate is reduced by a factor of 10 for the final $1000/2000$ iterations, respectively.

Byzantine Attacks. Next, we describe the attacks that the Byzantine workers employ in our experiments. Let $\bar{g}$ denote the average of honest messages in some round. We consider the following attacks:

- **Sign-flip (SF, Allen-Zhu et al., 2020):** each Byzantine worker computes a stochastic gradient and returns its negative.
- **Inner-Product Manipulation (IPM, Xie et al., 2020):** all Byzantine workers return the negative, re-scaled average of the honest messages, i.e., $-\epsilon\bar{g}$. We follow Karimireddy et al. (2021) and use $\epsilon = 0.1$.
- **A Little is Enough (ALIE, Baruch et al., 2019):** Let σ denote the (element-wise) standard deviation vector of honest messages. All Byzantine workers return the vector $\bar{g} - z\sigma$, where $z \in \mathbb{R}$ is computed as in Baruch et al. (2019); Karimireddy et al. (2021) (cf. Appendix G in the latter). For the MNIST experiments with $m = 17$ and $\delta m = 8$, we have $z \approx 1.22$.

J.1. MNIST Classification with Periodic Identity-Switching Strategy

In Figure 5, we show the full training curves on MNIST, corresponding to the final accuracy results presented in Figure 1. That is, we show the test accuracy as a function of the number of observed samples ($\times$ batch size $\times$ # of workers, i.e., total sample complexity) under the **Periodic** switching strategy for different values of K, where Byzantine workers employ the SF attack and the server uses CWTM. For the momentum and SGD methods, we used an initial learning rate of $\eta = 0.01$ and for DynaBRO it is $\eta = 0.05$. We observe that our method performs well across all values of K. However, when the switch rate K is small (specifically, smaller than the effective momentum horizon of $\frac{1}{1-\beta}$), momentum SGD fails to learn something meaningful and performs similarly to a random guess. When K is larger, momentum performs better, e.g., when $K = 100$ momentum with parameter $\beta = 0.9$ slightly outperforms our method.

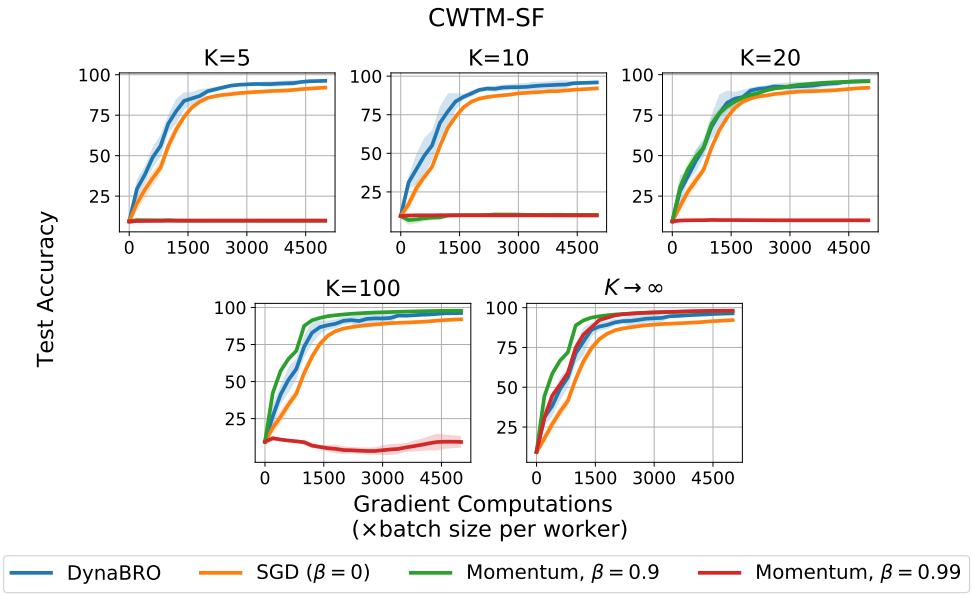

Figure 5. Test accuracy on MNIST under the **Periodic**(K) identity-switching strategy for different values of K. Byzantine workers employ SF attack and the server implements CWTM aggregation.

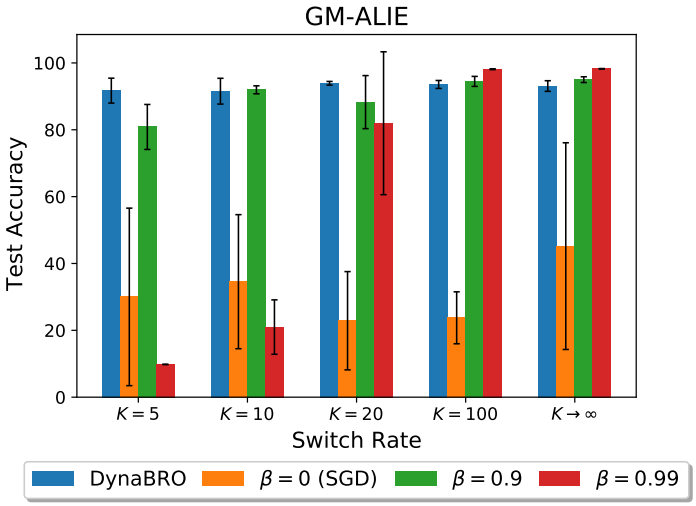

Figure 6. Final test accuracy on MNIST under the **Periodic**(K) identity-switching strategy for different values of K. Byzantine workers employ ALIE attack and the server implements geometric median (GM) aggregation.

In Figures 6 and 7, we show the test accuracy on the same MNIST configuration, but with Byzantine workers utilizing the ALIE attack (Baruch et al., 2019) and the server using the geometric median (GM, Pillutla et al., 2022) aggregator. For this

configuration, all methods are trained with the same initial learning rate of $\eta = 0.01$. While here momentum with parameter 0.9 performs reasonably when $K = 5$, the general trend is similar – our method has consistent accuracy across different values of K and momentum improves as K increases.

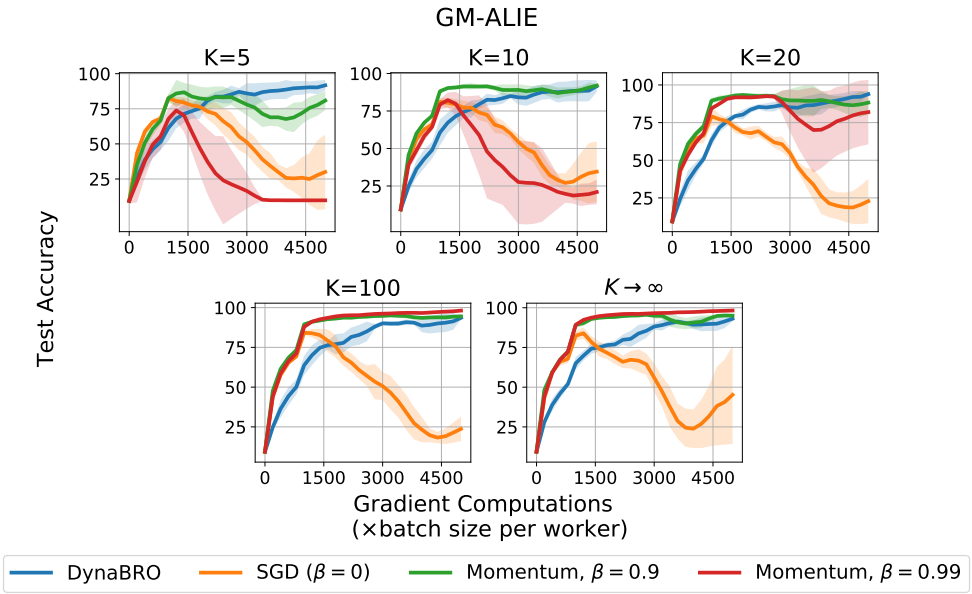

Figure 7. Test accuracy curves corresponding to the final results presented in Figure 6.

J.2. CIFAR-10 Classification with Bernoulli Identity-Switching Strategy

In this section, we present an additional configuration for the CIFAR-10 task under the **Bernoulli**$(p, D, \delta_{\max})$ switching strategy, with a different maximum fraction of Byzantine workers. Instead of using $\delta_{\max} = 0.72$ (maximum 18 Byzantine workers per iteration), we limit the maximum to 12 Byzantine workers, corresponding to $\delta_{\max} = 0.48$. This demonstrates that similar results to those in the main text are observed when there are fewer than half Byzantine workers in all iterations. For both configurations, we used an initial learning rate of $\eta = 0.01$ for momentum and SGD, and $\eta = 0.05$ for DynaBRO. In Figure 8, we show the results when $\delta_{\max} = 0.48$, mirroring those in Figure 2.

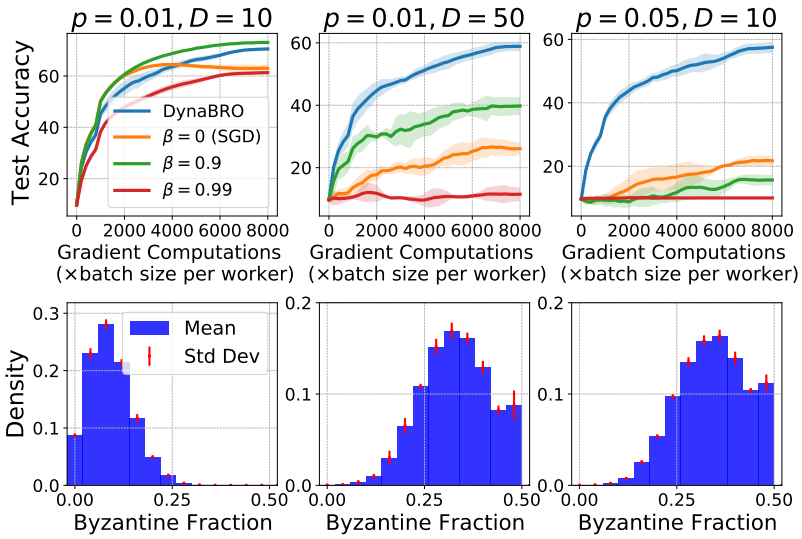

Figure 8. Test accuracy and histogram of the fraction of Byzantine workers on CIFAR-10 under the **Bernoulli**$(p, D, \delta_{\max})$ strategy for different values of p and D with $\delta_{\max} = 0.48$. Byzantine workers employ the IPM attack and the server uses CWMed.