Error Feedback for Muon and Friends

Kaja Gruntkowska KAJA.GRUNTKOWSKA@KAUST.EDU.SA
Alexander Gaponov ALIAKSANDR.HAPONAU@KAUST.EDU.SA
Zhirayr Tovmasyan ZHIRAYR.TOVMASYAN@KAUST.EDU.SA
Peter Richtárik PETER.RICHTARIK@KAUST.EDU.SA
King Abdullah University of Science and Technology, Saudi Arabia

Abstract

Recent optimizers like Muon, Scion, and Gluon have pushed the frontier of large-scale deep learning by exploiting layer-wise linear minimization oracles (LMOs) over non-Euclidean norm balls, capturing neural network structure in ways traditional algorithms cannot. Yet, no principled distributed framework exists, and communication bottlenecks remain unaddressed. Existing solutions are largely heuristic and lack any theoretical support. We introduce EF21-Muon, the first communication-efficient, non-Euclidean LMO-based optimizer with rigorous convergence guarantees. EF21-Muon supports stochastic gradients, momentum, and bidirectional compression with error feedback, recovering Muon/Scion when compression is off and specific norms are chosen–providing the first efficient distributed implementation of this powerful family. Our theory covers non-Euclidean layer-wise smooth and the sharper layer-wise (L^0, L^1)–smooth setting, matching best-known Euclidean rates while enabling faster convergence under suitable norms. Experiments on language modeling tasks confirm that EF21-Muon delivers significant communication savings without accuracy loss.

1. Introduction

Over the past decade, Adam and its variants [30, 41] have established themselves as the cornerstone of optimization in deep learning, owing to their empirical success across a wide range of tasks. Yet growing evidence suggests their dominance may be giving way to a new class of optimizers better suited to the geometry and scale of modern deep networks. Leading this shift are Muon [24] and methods inspired by it–Scion [53] and Gluon [61]–which replace Adam's global moment estimation with layer-wise, geometry-aware updates. Central to their design are layer-specific *linear minimization oracles* (LMOs) over non-Euclidean norm balls, enabling better alignment with the anisotropic structure of neural loss landscapes. Though relatively new, these optimizers are gaining traction, supported by a growing body of theoretical insights, community adoption, and strong empirical results, especially in training large language models (LLMs) [38, 44, 53, 67, 75, 77]. Despite this momentum, their development remains less mature than that of more established methods. Significant gaps persist–both in theory and practice–that must be addressed to fully realize their potential and make them truly competitive for the demands of ultra-scale learning.

Scaling Up. Modern machine learning (ML) thrives on scale. Today's state-of-the-art models are powered by elaborate architectures trained on massive datasets, often requiring weeks or months of computation [8, 76]. Naturally, this brings new demands on optimization: training on a single machine is no longer viable [11, 85], making distributed computing the default. Mathematically, this task can be modeled as the (generally non-convex) optimization problem

$$\min_{X \in \mathcal{S}} \left\{ f(X) := \tfrac{1}{n} \textstyle\sum_{j=1}^{n} f_j(X) \right\}, \tag{1}$$

where $X \in \mathcal{S}$ represents the model parameters, $n \geq 1$ is the number of workers/clients/machines, and $f_j(X)$ is the local loss of the model X on the data stored on worker $j \in [n] := \{1, \ldots, n\}$. We consider the general heterogeneous setting, where the local objectives f_j may differ arbitrarily across machines, reflecting real-world scenarios such as multi-datacenter pipelines or federated learning [32, 43]. Here, $\mathcal{S}$ is a d-dimensional vector space equipped with an inner product $\langle \cdot, \cdot \rangle : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$, inducing the standard Euclidean norm $\|\cdot\|_2$. Furthermore, we endow $\mathcal{S}$ with an arbitrary norm $\|\cdot\| : \mathcal{S} \to \mathbb{R}_{\geq 0}$. The corresponding dual norm $\|\cdot\|_\star : \mathcal{S} \to \mathbb{R}_{\geq 0}$ is defined via $\|X\|_\star := \sup_{\|Z\| \leq 1} \langle X, Z \rangle$. The general algorithmic framework introduced in this work gives rise to a variety of methods arising from different norm choices. In matrix spaces, a particularly important class is the family of *operator norms*, defined by $\|A\|_{\alpha \to \beta} := \sup_{\|Z\|_\alpha = 1} \|AZ\|_\beta$.

Communication: the Cost of Scale. In client-server architectures, coordination is centralized, with workers performing local computations and periodically synchronizing with the coordinator [80, 81]. While this design unlocks learning at unprecedented scales, it introduces a critical bottleneck: *communication*. Modern models' size places a heavy burden on channels used to synchronize updates, as each step may require transmitting large d-dimensional vectors (e.g., parameters or gradients) over links often slower than local computation [25, 37, 50, 82]. Without communication-efficient strategies, this imbalance dominates costs and limits distributed optimization efficiency.

Distributed Muon: Bridging the Gap. The case for communication-efficient distributed training is clear, as is the promise of Muon for deep learning. The natural question is: can we merge the two? This intersection remains largely unexplored. Nonetheless, three recent efforts are worth noting. Liu et al. [38] propose a distributed variant of Muon based on ZeRO-1 [58], Thérien et al. [75] show that Muon can be used instead of AdamW as the inner optimizer in DiLoCo, and Ahn et al. [1] introduce Dion, a Muon-inspired algorithm compatible with 3D parallelism that employs low-rank approximations for efficient orthonormalized updates. While promising empirically, these methods *lack any formal theoretical guarantees*. Our goal is to close this gap by developing a distributed optimizer leveraging non-Euclidean geometry that not only works in practice but also comes with strong provable convergence guarantees. Our central question is:

Can we efficiently distribute Muon *without compromising its theoretical and practical benefits?*

We provide an affirmative answer through the following **contributions**:

1. **Compressed non-Euclidean distributed optimization.** We propose EF21-Muon, an LMO-based distributed algorithm with bidirectionally compressed updates with error feedback. Parameterized by the norm in the LMO step, EF21-Muon recovers a broad class of compressed methods, and, for spectral norms, yields *the first communication-efficient distributed variants of* Muon *and* Scion.

2. **Practical deep learning variant.** The main text presents a simplified global version (Algorithm 1), while our main algorithms are designed and analyzed in a *layer-wise* manner (Algorithms 2 and 3), reflecting the structure of neural networks. This allows us to better align with practice (methods like Muon are applied *per-layer*) and to introduce *anisotropic modeling assumptions*.

Algorithm 1 EF21-Muon (simplified)

```
Input: radii $t^k > 0$; momentum parameter $\beta \in (0, 1]$; initial iterate $X^0 \in \mathcal{S}$ (stored on the server);
       initial iterate shift $W^0 = X^0$ (stored on the server and the workers); initial gradient estimators
       $G_j^0$ (stored on the workers); $G^0 = \frac{1}{n}\sum_{j=1}^n G_j^0$ (stored on the server); initial momentum
       $M_j^0$ (stored on the workers); worker compressors $\mathcal{C}_j^k$; server compressors $\mathcal{C}^k$
for $k = 0, 1, \ldots, K-1$ do
    $X^{k+1} = \mathrm{LMO}_{\mathcal{B}(X^k, t^k)}\left(G^k\right)$                    // Take LMO-type step
    $S^k = \mathcal{C}^k(X^{k+1} - W^k)$                    // Compress shifted model on the server
    $W^{k+1} = W^k + S^k$                    // Update model shift
    Broadcast $S^k$ to all workers
    for $j = 1, \ldots, n$ in parallel do
        $W^{k+1} = W^k + S^k$                    // Update model shift
        $M_j^{k+1} = (1-\beta)M_j^k + \beta\nabla f_j(W^{k+1}, \xi_j^{k+1})$                    // Compute momentum
        $R_j^{k+1} = \mathcal{C}_j^k(M_j^{k+1} - G_j^k)$                    // Compress shifted gradient
        $G_j^{k+1} = G_j^k + R_j^{k+1}$
        Broadcast $R_j^{k+1}$ to the server
    end
    $G^{k+1} = \frac{1}{n}\sum_{j=1}^n G_j^{k+1} = G^k + \frac{1}{n}\sum_{j=1}^n R_j^{k+1}$                    // Compute gradient estimator
end
```

3. **Strong convergence guarantees.** EF21-Muon enjoys strong theoretical guarantees under non-Euclidean smoothness (Theorems 3 and 5) and non-Euclidean (L^0, L^1)–smoothness (Theorems 4 and 6), matching state-of-the-art Euclidean rates and potentially improving convergence under well-chosen norms. These results are subsumed by our more general analysis of the layer-wise methods under *layer-wise non-Euclidean smoothness* (Theorems 25 and 30) and *layer-wise non-Euclidean* (L^0, L^1)*–smoothness* (Theorems 28 and 36).

5. **Strong empirical performance.** Experiments with `NanoGPT`-124M on `FineWeb` confirm that EF21-Muon can significantly reduce communication while preserving accuracy (Sections 5 and 5).

2. Background

We frame problem (1) in an abstract vector space $\mathcal{S}$. In several of our results, the specific structure of $\mathcal{S}$ does not matter. One may view $\mathcal{S}$ as simply $\mathbb{R}^d$, in which case the model parameters are flattened into a $d \times 1$ vector. However, in the context of deep learning, it is often useful to explicitly model the layer-wise structure (see Section C). In such cases, $X \in \mathcal{S}$ represents the collection of matrices $X_i \in \mathcal{S}_i := \mathbb{R}^{m_i \times n_i}$ of trainable parameters across all layers $i = 1, \ldots, p$ of the network with a total number $d := \sum_{i=1}^p m_i n_i$ of parameters. Then, $\mathcal{S}$ is the d-dimensional product space $\mathcal{S} := \bigotimes_{i=1}^p \mathcal{S}_i \equiv \mathcal{S}_1 \otimes \cdots \otimes \mathcal{S}_p$, and we write $X = [X_1, \ldots, X_p]$.

What is Muon**?** Muon, introduced by Jordan et al. [24], is an optimizer for the hidden layers of neural networks.[1] For clarity, assume that X represents a single layer of the network (a full

[1] The first and last layers are optimized using other optimizers–see Section C.1 for details.

layer-wise description is provided in Section C.1). In this setting, Muon updates $X^{k+1} = X^k - t^k U^k (V^k)^\top$, where $t^k > 0$ and the matrices U^k, V^k are derived from the SVD of the momentum matrix $G^k = U^k \Sigma^k (V^k)^\top$. This update is a special case of the *linear minimization oracle* (LMO) framework

$$X^{k+1} = X^k + t^k \mathrm{LMO}_{\mathcal{B}(0,1)}\left(G^k\right), \tag{2}$$

where $\mathcal{B}(X,t) := \{Z \in \mathcal{S} : \|Z - X\| \leq t\}$ and $\mathrm{LMO}_{\mathcal{B}(X,t)}(G) := \arg\min_{Z \in \mathcal{B}(X,t)} \langle G, Z \rangle$. Muon corresponds to the case where $\|\cdot\| = \|\cdot\|_{2\to 2}$ is the spectral norm, in which case the LMO reduces to $\mathrm{LMO}_{\mathcal{B}(0,1)}\left(G^k\right) = -U^k (V^k)^T$. Recent analyses [33, 53, 61] focus on the general update (2). Among them, Pethick et al. [53] introduce Scion, which extends Muon to all layers, and Riabinin et al. [61] develop Gluon–a general LMO-based framework with stronger convergence guarantees encompassing Muon and Scion as special cases. We adopt this unifying viewpoint by treating all three algorithms as instances of Gluon, using it as the umbrella term for the entire class.

Distributing the LMO. Distributing (2) is far from trivial, as the limited literature suggests. To illustrate the difficulty, consider a deterministic version of (2), where the momentum term G^k is replaced by the exact gradient $\nabla f(X^k)$. Applied to problem (1), the iteration becomes

$$X^{k+1} = X^k + \mathrm{LMO}_{\mathcal{B}(0,t^k)}\left(\tfrac{1}{n}\textstyle\sum_{j=1}^n \nabla f_j(X^k)\right).$$

The most basic approach to distributing this update consists of the following four main steps:

1. Each worker computes its local gradient $\nabla f_j(X^k)$ at iteration k.
2. **w2s:** The workers send their gradients $\nabla f_j(X^k)$ to the central server.
3. The server averages these gradients and computes the LMO update.
4. **s2w:** The server sends X^{k+1} (or $\mathrm{LMO}_{\mathcal{B}(0,t^k)}(\cdot)$) back to the workers.

This scheme involves two potentially costly phases: (1) workers-to-server (= w2s) and (2) server-to-workers (= s2w) communication. As each transmitted object resides in $\mathcal{S}$, every iteration involves exchanging dense, d-dimensional data, imposing substantial communication overhead that can quickly overwhelm available resources. This is precisely where compression comes into play.

Compression. Compression has been the subject of extensive study in the Euclidean regime [2, 19, 63]. It is typically achieved by applying a (potentially randomized) operator $\mathcal{C}$ that maps the message X to a more compact, cheaper to transmit representation $\mathcal{C}(X)$. The literature distinguishes between two main classes of compression operators: *unbiased* and *biased* (or *contractive*) compressors. We focus on the latter class, which is known to perform better empirically [5, 66].

Definition 1 (Contractive compressor) *A (possibly randomized) mapping $\mathcal{C} : \mathcal{S} \to \mathcal{S}$ is a contractive compression operator with parameter $\alpha \in (0, 1]$ if*

$$\mathbb{E}\left[\|\mathcal{C}(X) - X\|^2\right] \leq (1 - \alpha)\|X\|^2 \qquad \forall X \in \mathcal{S}. \tag{3}$$

Remark 2 *The classical definition of a contractive compressor is based on the Euclidean norm, i.e., $\|\cdot\| = \|\cdot\|_2$ in (3). Here, condition (3) is expressed in terms of an arbitrary norm $\|\cdot\|$ for greater flexibility. Section E gives several examples satisfying this generalized definition. Depending on the compression objective, we will assume that (3) holds with respect to $\|\cdot\|$, its dual $\|\cdot\|_\star$, or $\|\cdot\|_2$, and denote the respective families of compressors as $\mathbb{B}(\alpha)$, $\mathbb{B}_\star(\alpha)$, and $\mathbb{B}_2(\alpha)$.*

3. Non-Euclidean Distributed Training

We develop the *first communication-efficient variant of* Gluon (and by extension, its special cases Muon and Scion), called EF21-Muon, which combines biased compression, gradient stochasticity, and momentum while retaining *strong theoretical guarantees*. Its simplified version, applied globally to the full parameter vector X, is shown in Algorithm 1. A more general, deep learning–oriented layer-wise variant operating in the product space $\mathcal{S} := \bigotimes_{i=1}^{p} \mathbb{R}^{m_i \times n_i}$ is given in Algorithm 3. For clarity, we focus on the global version in the main text, noting that all results are special cases of the layer-wise guarantees in Section F. While the pseudocode is largely self-explanatory (for a more detailed description, see Section C.2), we highlight the most important components:

◇ **Role of Compression.** Compression is key for reducing communication overhead in distributed training. Algorithm 1 adheres to this principle by transmitting only the compressed messages S^k and R_j^k, never the full dense updates. When compression is disabled (i.e., $\mathcal{C}_j^k$ and $\mathcal{C}^k$ are identity mappings) and $n = 1$, EF21-Muon reduces exactly to Gluon, which in turn recovers Muon and Scion.

◇ **Role of Error Feedback.** Even in the Euclidean setup, biased compression can break distributed Gradient Descent (GD) unless Error Feedback (EF) is used (see Section B.2). To remedy this, we adopt a modern EF strategy inspired by EF21 [63] for the w2s direction. Its role is to stabilize training and prevent divergence. To reduce s2w communication overhead, we further incorporate the primal compression mechanism of EF21-P [17].

◇ **Role of Gradient Stochasticity.** In large-scale ML, computing full gradients $\nabla f_j(x)$ is typically computationally infeasible. In practice, they are replaced with stochastic estimates, which drastically reduces per-step computational cost and makes the method scalable to practical workloads.

◇ **Role of Momentum.** Stochastic gradients inevitably introduce noise into the optimization process. Without further stabilization, this leads to high variance and convergence only to a neighborhood of the solution. Momentum plays a critical role in mitigating this issue: it reduces variance in the updates, accelerates convergence, and eliminates the persistent oscillations.

4. Convergence Results

We adopt standard lower-boundedness assumptions on f, and in certain cases, also on f_j, $j \in [n]$.

Assumption 1 *There exist $f^\star \in \mathbb{R}$ such that $f(X) \geq f^\star$ for all $X \in \mathcal{S}$.*

Assumption 2 *For all $j \in [n]$, there exist $f_j^\star \in \mathbb{R}$ such that $f_j(X) \geq f_j^\star$ for all $X \in \mathcal{S}$.*

We study two smoothness regimes. The first, standard L–smoothness generalized to arbitrary norms, is the default in virtually all convergence results for Muon and Scion [33, 36, 53].

Assumption 3 *The function f is L–smooth, i.e., $\left\|\nabla f(X) - \nabla f(Y)\right\|_\star \leq L \left\|X - Y\right\|$ for all $X, Y \in \mathcal{S}$. Moreover, the functions f_j are L_j–smooth for all $j \in [n]$. We define $\tilde{L}^2 := \frac{1}{n}\sum_{j=1}^{n} L_j^2$.*

To our knowledge, the only exception departing from this standard setting is the recent work on Gluon [61]. The authors argue that layer-wise optimizers are designed specifically for deep learning, where classical smoothness fails [87]. Instead, they build upon the (L^0, L^1)–smoothness model

[87],[2] a strictly weaker alternative motivated by empirical observations from NLP training dynamics. Riabinin et al. [61] introduce a *layer-wise* variant (Assumption 8), arguing that the heterogeneity across layers requires smoothness constants to vary accordingly. Consistent with this line of work, we provide guarantees under the general *layer-wise* (L^0, L^1)*–smoothness* assumption in Appendices F.3.2 and F.4.2. For ease of exposition, the main text treats the case of a generic vector space, where the assumption reduces to a non-Euclidean variant of asymmetric (L^0, L^1)–smoothness [7].

Assumption 4 *The function $f : \mathcal{S} \mapsto \mathbb{R}$ is (L^0, L^1)–smooth, i.e., there exist $L^0, L^1 > 0$ such that $\|\nabla f(X) - \nabla f(Y)\|_\star \leq \left(L^0 + L^1 \|\nabla f(X)\|_\star\right) \|X - Y\|$ for all $X, Y \in \mathcal{S}$. Moreover, the functions f_j, $j \in [n]$, are (L^0_j, L^1_j)–smooth. We define $L^1_{\max} := \max_{j\in[n]} L^1_j$ and $\bar{L}^0 := \frac{1}{n}\sum_{j=1}^n L^0_j$.*

Assumption 4 is strictly more general than Assumption 3, as it allows the smoothness constant to grow with the norm of the gradient, a key property observed in deep learning [87].

Deterministic setting. As a warm-up, we first present the convergence guarantees of the deterministic counterpart of Algorithm 1 (formalized in Algorithm 2), starting with the smooth setting.

Theorem 3 *Let Assumptions 1 and 3 hold. Let $\{X^k\}_{k=0}^{K-1}$, $K \geq 1$, be the iterates of Algorithm 2 (with $p = 1$) initialized with $X^0 = W^0$, $G^0_j = \nabla f_j(X^0)$, $j \in [n]$, and run with $\mathcal{C}^k \in \mathbb{B}(\alpha_P)$, $\mathcal{C}^k_j \in \mathbb{B}_\star(\alpha_D)$ and $0 < \gamma^k \equiv \gamma \leq \left(2L + {}^4\!/\!_{\alpha_D}\sqrt{12 + {}^{66}\!/\!_{\alpha_P^2}}\tilde{L}\right)^{-1}$. Then*

$$\tfrac{1}{K}\textstyle\sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\nabla f(X^k)\right\|_\star^2\right] \leq \tfrac{4\left(f(X^0)-f^\star\right)}{K\gamma}.$$

Theorem 3 is a special case of the layer-wise result in Theorem 25. To our knowledge, no prior work analyzes comparable compressed methods under general non-Euclidean geometry. In the Euclidean case, our guarantees recover known results: without primal compression ($\alpha_P = 1$), they match the rate of Richtárik et al. [63, Theorem 1]; with primal compression, they align with the rate of EF21-BC from Fatkhullin et al. [12, Theorem 21] (however, Algorithm 2 and EF21-BC differ algorithmically, and the former does not reduce to the latter in the Euclidean setting).

In the generalized smooth setting, we prove convergence without primal compression. As discussed in Section E.1, s2w communication can still be efficient in the non-Euclidean regime, since LMOs under certain norms inherently exhibit *compression-like behavior*.

Theorem 4 *Let Assumptions 1, 2 and 4 hold and let $\{X^k\}_{k=0}^{K-1}$, $K \geq 1$, be the iterates of Algorithm 2 (with $p = 1$) initialized with $G^0_j = \nabla f_j(X^0)$, $j \in [n]$, and run with $\mathcal{C}^k \equiv \mathcal{I}$ (the identity compressor), $\mathcal{C}^k_j \in \mathbb{B}_\star(\alpha_D)$, and $t^k \equiv {}^{\eta}\!/\!_{\sqrt{K+1}}$ for some $\eta > 0$. Then,*

$$\min_{k=0,\dots,K} \mathbb{E}\left[\left\|\nabla f(X^k)\right\|_\star\right] \leq \frac{\exp\left(4\eta^2 C L^1_{\max}\right)}{\eta\sqrt{K+1}}\delta^0 + \frac{\eta\left(4C\frac{1}{n}\sum\limits_{j=1}^n L^1_j\left(f^\star - f^\star_j\right) + C\frac{1}{n}\sum\limits_{j=1}^n \frac{L^0_j}{L^1_j} + D\right)}{\sqrt{K+1}},$$

where $C := \frac{L^1}{2} + \frac{2\sqrt{1-\alpha_D}L^1_{\max}}{1-\sqrt{1-\alpha_D}}$ and $D := \frac{L^0}{2} + \frac{2\sqrt{1-\alpha_D}\bar{L}^0}{1-\sqrt{1-\alpha_D}}$.

Theorem 4 (a corollary of Theorem 28), similar to Theorem 3, establishes the desirable $\mathcal{O}({}^1\!/\!_{\sqrt{K}})$ rate for expected gradient norms. Unlike the stepsizes in Theorem 3, the radii here are *independent of problem-specific constants* (though known smoothness can inform the choice of η). In the Euclidean case, our result matches the rate of ‖EF21‖ under (L^0, L^1)–smoothness [29].

[2] The original (L^0, L^1)–smoothness assumption of Zhang et al. [87] was defined for twice-differentiable functions via Hessian norms. This notion and our formulation in Assumption 4 are closely related–see Chen et al. [7]

Stochastic setting. We now turn to the convergence guarantees of our practical variant of EF21-Muon (Algorithm 1), which incorporates noisy gradients and momentum. We assume access to a standard stochastic gradient oracle $\nabla f(\cdot, \xi)$, $\xi \sim \mathcal{D}$ with bounded variance.

Assumption 5 *The stochastic gradient estimator $\nabla f(\cdot, \xi) : \mathcal{S} \mapsto \mathcal{S}$ is unbiased and has bounded variance. That is, $\mathbb{E}_{\xi \sim \mathcal{D}}[\nabla f(X, \xi)] = \nabla f(X)$ for all $X \in \mathcal{S}$ and there exists $\sigma \geq 0$ such that $\mathbb{E}_{\xi \sim \mathcal{D}}\left[\|\nabla f_j(X, \xi) - \nabla f_j(X)\|_2^2\right] \leq \sigma^2$ for all $X \in \mathcal{S}$.*

Assumption 5 uses the Euclidean norm to facilitate the bias-variance decomposition. Since $\mathcal{S}$ is finite-dimensional, the magnitudes measured in $\|\cdot\|_2$ can be related to quantities measured in $\|\cdot\|$ via norm equivalence: there exist $\underline{\rho}, \bar{\rho} > 0$ such that $\underline{\rho}\,\|X\| \leq \|X\|_2 \leq \bar{\rho}\,\|X\|$ for all $X \in \mathcal{S}$.
We first analyze the smooth case; the theorem below is a special case of Theorem 30.

Theorem 5 *Let Assumptions 1, 3 and 5 hold. Let $\{X^k\}_{k=0}^{K-1}$, $K \geq 1$, be the iterates of Algorithm 1 initialized with $X^0 = W^0$, $G_j^0 = M_j^0 = \nabla f_j(X^0)$, $j \in [n]$, and run with $\mathcal{C}^k \in \mathbb{B}(\alpha_P)$, $\mathcal{C}_j^k \in \mathbb{B}_2(\alpha_D)$, any $\beta \in (0, 1]$, and $0 \leq \gamma^k \equiv \gamma \leq \left(2\sqrt{\zeta} + 2L\right)^{-1}$, where $\zeta := \frac{\bar{\rho}^2}{\underline{\rho}^2}\left(\frac{12}{\beta^2}L^2 + \frac{24(\beta+2)}{\alpha_P^2}L^2 + \frac{36\left(\beta^2+4\right)}{\alpha_D^2}\tilde{L}^2 + \frac{144\beta^2(2\beta+5)}{\alpha_P^2\alpha_D^2}\tilde{L}^2\right)$. Then*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(X^k)\right\|_\star^2\right] \leq \frac{4\left(f(X^0)-f^\star\right)}{K\gamma} + 24\left(\frac{1}{n} + \frac{(1-\alpha_D)\beta}{\alpha_D} + \frac{12\beta^2}{\alpha_D^2}\right)\sigma^2\bar{\rho}^2\beta.$$

With $\gamma^k \equiv \left(2\sqrt{\zeta} + 2L\right)^{-1}$ and $\beta = \min\left\{1, \left(\frac{\delta^0 L^0 n}{\underline{\rho}^2\sigma^2 K}\right)^{1/2}, \left(\frac{\delta^0 L^0 \alpha_D}{\underline{\rho}^2\sigma^2 K}\right)^{1/3}, \left(\frac{\delta^0 L_1^0 \alpha_D^2}{\underline{\rho}^2\sigma^2 K}\right)^{1/4}\right\}$, where $\delta^0 := f(X^0) - f^\star$, the result guarantees $\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(X^k)\right\|_\star^2\right] = \mathcal{O}(1/\sqrt{K})$ (Theorem 31). In the Euclidean case without primal compression ($\bar{\rho}_i^2 = \underline{\rho}_i^2 = 1$, $\alpha_P = 1$), Theorem 5 matches the convergence rate of EF21-SDGM established in Fatkhullin et al. [13, Theorem 3] (Theorem 33).

Theorem 6 *Let Assumptions 1, 2, 4 and 5 hold. Let $\{X^k\}_{k=0}^{K-1}$, $K \geq 1$, be the iterates of Algorithm 1 initialized with $M_j^0 = \nabla f_j(X^0; \xi_j^0)$, $G_j^0 = \mathcal{C}_j^0(\nabla f_j(X^0; \xi_j^0))$, $j \in [n]$, and run with $\mathcal{C}^k \equiv \mathcal{I}$ (the identity compressor), $\mathcal{C}_j^k \in \mathbb{B}_2(\alpha_D)$, $\beta = 1/(K+1)^{1/2}$ and $0 \leq t^k \equiv t = \eta/(K+1)^{3/4}$, where $\eta^2 \leq \min\left\{\frac{(K+1)^{1/2}}{6(L^1)^2}, \frac{(1-\sqrt{1-\alpha_D})\underline{\rho}(K+1)^{1/2}}{24\sqrt{1-\alpha_D}\bar{\rho}(L_{\max}^1)^2}, \frac{\beta\underline{\rho}(K+1)^{1/2}}{24\bar{\rho}(L_{\max}^1)^2}, 1\right\}$. Then*

$$\min_{k=0,\ldots,K}\mathbb{E}\left[\left\|\nabla f(X^k)\right\|_\star\right] \leq \frac{3\left(f(X^0)-f^\star\right)}{\eta(K+1)^{1/4}} + \frac{\eta L^0}{(K+1)^{3/4}} + \frac{16\sqrt{1-\alpha_D}\bar{\rho}\sigma}{(1-\sqrt{1-\alpha_D})(K+1)^{1/2}} + \frac{8\bar{\rho}\sigma}{\sqrt{n}(K+1)^{1/4}}$$
$$+\frac{\eta\bar{\rho}}{\underline{\rho}}\left(\frac{8}{(K+1)^{1/4}} + \frac{8\sqrt{1-\alpha_D}}{(1-\sqrt{1-\alpha_D})(K+1)^{3/4}}\right)\left(\frac{1}{n}\sum_{j=1}^n(L_j^1)^2\left(f^\star - f_j^\star\right) + \bar{L}^0\right).$$

Theorem 6 (a corollary of Theorem 36) establishes an $\mathcal{O}(1/K^{1/4})$ convergence rate, matching the state-of-the-art for SGD-type methods applied to non-convex functions [10, 72]. In the Euclidean setting, it recovers the rate of ||EF21-SDGM|| established in Khirirat et al. [29, Theorem 2].

5. Experimental Highlights

We present key experimental results below, with additional details and extended experiments available in Section G.[3]

[3] Code for experiments is available here.

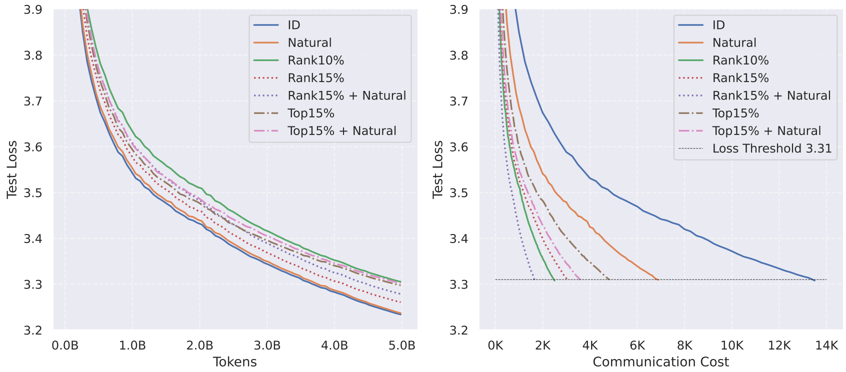

Figure 1: Left: Test loss vs. # of tokens processed. Right: Test loss vs. # of bytes sent from each worker to the server normalized by model size to reach test loss 3.31. Rank/Top$X\%$ = Rank/Top$K$ compressor with sparsification level $X\%$; ID = no compression.

Experimental setup. All experiments are conducted on 4 NVIDIA Tesla V100-SXM2-32GB GPUs or 4 NVIDIA A100-SXM4-80GB in a Distributed Data Parallel (DDP) setup. The dataset is evenly partitioned across workers, with one worker node acting as the master, aggregating compressed updates from the others. Training and evaluation are implemented in PyTorch,[4] extending open-source codebases [27, 52, 60].

We train a `NanoGPT` model [27] with 124M parameters on the `FineWeb10B` dataset [51], using input sequences of length 1024 and a batch size of 256. Optimization is performed with EF21-Muon, using spectral norm LMOs for hidden layers and ℓ_∞ norm LMOs for embedding and output layers (which coincide due to weight sharing), following the approach of Pethick et al. [53]. For spectral norm LMOs, inexact updates are computed with 5 Newton–Schulz iterations [6, 34], as in Jordan et al. [24].

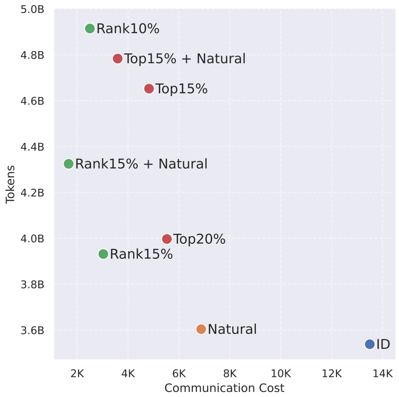

Figure 2: Trade-off between token efficiency and communication cost for different compression setups at a target test loss of 3.31.

Following common practice in communication compression literature, we assume that broadcasting is free and focus on w2s communication. Thus, the server-side compressor is fixed to $\mathcal{I}$, while worker compressors vary among Top$K$, Rank$K$ [65], Natural compressor [19] and combinations thereof: Top$K$ + Natural compressor of selected elements, and Rank$K$ + Natural compressor applied to all components of the low rank decomposition. These are tested under multiple compression levels and compared against an uncompressed baseline (i.e., standard Scion/Gluon; see Section 3). Learning rates are tuned per optimizer and experimental setting, initialized from the values in the Gluon repository [60] (see Section G.3). We adopt the same learning rate scheduler as Karpathy [27] and fix the momentum parameter to 0.9. Model and optimizer hyperparameters are summarized in Tables 2 and 4, respectively.

[4] PyTorch Documentation: https://pytorch.org/docs/stable/index.html

Results. For Rank$K$ and Top$K$ compressors, we evaluate multiple compression levels (in plots, Rank$X\%$/Top$X\%$ denotes a Rank$K$/Top$K$ compressor with compression level $X\%$). We report experimental results for a 5B-token training budget ($> 40\times$ model size) in Figure 1 (left), and to reach a strong loss threshold of 3.31 in Figures 1 (right) and 2.

Table 1: Communication cost per round (in bytes), normalized relative to the identity compressor.

Compressor	Relative Cost
ID	1.0000
Natural	0.5000
Rank20%	0.2687
Rank15%	0.2019
Rank15% + Natural	0.1010
Rank10%	0.1335
Rank10% + Natural	0.0667
Rank5%	0.0667
Top20%	0.3625
Top15%	0.2718
Top15% + Natural	0.1969
Top10%	0.1812
Top10% + Natural	0.1312
Top5%	0.0906

The number of tokens required to reach a target loss depends on the compressor. Figure 2 provides a comparison of the numbers of tokens used in the training run to reach a strong test loss threshold of 3.31 plotted against the communication cost (reported as the number of bits transmitted from each worker to the server normalized by the model size), plotted against the w2s communication cost. Shorter 2.5B-token runs are reported in Section G.5 to assess performance under limited training budgets.
In Figure 1, we plot test loss vs. tokens processed, as well as the w2s communication cost required to reach the 3.31 loss threshold. For each compressor, we report its most competitive configuration (see Section G.4 for a detailed ablation). As expected, compression slows convergence in terms of number of training steps, but substantially reduces per-step communication cost (Table 1). Overall, this yields significant **communication savings—up to** $7\times$ for Rank15% + Natural compressor, and roughly $4\times$ for Top15% + Natural compressor—relative to the uncompressed baseline.

References

[1] Kwangjun Ahn, Byron Xu, Natalie Abreu, and John Langford. Dion: Distributed orthonormalized updates. *arXiv preprint arXiv:2504.05295*, 2025. URL https://arxiv.org/abs/2504.05295.

[2] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in neural information processing systems*, 30, 2017. URL https://arxiv.org/abs/1610.02132.

[3] Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric Renggli. The convergence of sparsified gradient methods. *Advances in Neural Information Processing Systems*, 31, 2018. URL https://arxiv.org/abs/1809.10505.

[4] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569. PMLR, 2018. URL https://arxiv.org/abs/1802.04434.

[5] Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *arXiv preprint arXiv:2002.12410*, 2020. URL https://arxiv.org/abs/2002.12410.

[6] Å. Björck and C. Bowie. An iterative algorithm for computing the best estimate of an orthogonal matrix. *SIAM Journal on Numerical Analysis*, 8(2):358–364, 1971. URL https://doi.org/10.1137/0708036.

[7] Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *International Conference on Machine Learning*, pages 5396–5427. PMLR, 2023. URL https://arxiv.org/abs/2303.02854.

[8] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. URL https://arxiv.org/abs/2507.06261v4.

[9] Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signSGD. *Advances in neural information processing systems*, 35:9955–9968, 2022. URL https://arxiv.org/abs/2208.11195.

[10] Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *International conference on machine learning*, pages 2260–2268. PMLR, 2020. URL https://arxiv.org/abs/2002.03305.

[11] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc' aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, Quoc Le, and Andrew Ng. Large scale distributed deep networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/6aca97005c68f1206823815f66102863-Paper.pdf.

[12] Ilyas Fatkhullin, Igor Sokolov, Eduard Gorbunov, Zhize Li, and Peter Richtárik. EF21 with bells & whistles: Practical algorithmic extensions of modern error feedback. *arXiv preprint arXiv:2110.03294*, 2021. URL https://arxiv.org/abs/2110.03294.

[13] Ilyas Fatkhullin, Alexander Tyurin, and Peter Richtárik. Momentum provably improves error feedback! *Advances in Neural Information Processing Systems*, 36:76444–76495, 2023. URL https://arxiv.org/abs/2305.15155.

[14] Athanasios Glentis, Jiaxiang Li, Andi Han, and Mingyi Hong. A minimalist optimizer design for LLM pretraining. *arXiv preprint arXiv:2506.16659*, 2025. URL https://www.arxiv.org/abs/2506.16659.

[15] Eduard Gorbunov, Dmitry Kovalev, Dmitry Makarenko, and Peter Richtárik. Linearly converging error compensated SGD. *Advances in Neural Information Processing Systems*, 33:20889–20900, 2020. URL https://arxiv.org/abs/2010.12292.

[16] Eduard Gorbunov, Konstantin Burlachenko, Zhize Li, and Peter Richtárik. MARINA: Faster non-convex distributed learning with compression. In *38th International Conference on Machine Learning*, 2021. URL https://arxiv.org/abs/2102.07845.

[17] Kaja Gruntkowska, Alexander Tyurin, and Peter Richtárik. EF21-P and friends: Improved theoretical communication complexity for distributed optimization with bidirectional compression. In *International Conference on Machine Learning*, pages 11761–11807. PMLR, 2023. URL https://arxiv.org/pdf/2209.15218.

[18] Kaja Gruntkowska, Alexander Tyurin, and Peter Richtárik. Improving the worst-case bidirectional communication complexity for nonconvex distributed optimization under function similarity. *Advances in Neural Information Processing Systems*, 37:88807–88873, 2024. URL https://arxiv.org/abs/2402.06412.

[19] Samuel Horváth, Chen-Yu Ho, Ludovit Horvath, Atal Narayan Sahu, Marco Canini, and Peter Richtárik. Natural compression for distributed deep learning. In *Mathematical and Scientific Machine Learning*, pages 129–141. PMLR, 2022. URL https://arxiv.org/abs/1905.10988.

[20] Junxian Huang, Feng Qian, Alexandre Gerber, Z Morley Mao, Subhabrata Sen, and Oliver Spatscheck. A close examination of performance and power characteristics of 4g lte networks. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*, pages 225–238, 2012. URL https://dl.acm.org/doi/10.1145/2307636.2307658.

[21] Rustem Islamov, Xun Qian, and Peter Richtárik. Distributed second order methods with fast rates and compressed communication. In *International conference on machine learning*, pages 4617–4628. PMLR, 2021. URL https://arxiv.org/abs/2102.07158.

[22] Rustem Islamov, Xun Qian, Slavomír Hanzely, Mher Safaryan, and Peter Richtárik. Distributed newton-type methods with communication compression and bernoulli aggregation. *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2206.03588.

[23] Ruichen Jiang, Devyani Maladkar, and Aryan Mokhtari. Convergence analysis of adaptive gradient methods under refined smoothness and noise assumptions. *arXiv preprint arXiv:2406.04592*, 2024. URL https://arxiv.org/abs/2406.04592.

[24] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

[25] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021. URL https://arxiv.org/abs/1912.04977.

[26] Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes signSGD and other gradient compression schemes. In *International conference on*

machine learning, pages 3252–3261. PMLR, 2019. URL https://arxiv.org/abs/1901.09847.

[27] Andrej Karpathy. nanoGPT, 2023. URL https://github.com/karpathy/nanoGPT.

[28] Jonathan A Kelner, Yin Tat Lee, Lorenzo Orecchia, and Aaron Sidford. An almost-linear-time algorithm for approximate max flow in undirected graphs, and its multicommodity generalizations. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 217–226. SIAM, 2014. URL https://arxiv.org/abs/1304.2338.

[29] Sarit Khirirat, Abdurakhmon Sadiev, Artem Riabinin, Eduard Gorbunov, and Peter Richtárik. Error feedback under (l_0, l_1)-smoothness: Normalization and momentum. *arXiv preprint arXiv:2410.16871*, 2024. URL https://arxiv.org/abs/2410.16871.

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

[31] Anastasia Koloskova, Tao Lin, Sebastian U Stich, and Martin Jaggi. Decentralized deep learning with arbitrary communication compression. *arXiv preprint arXiv:1907.09356*, 2019. URL https://arxiv.org/abs/1907.09356.

[32] Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016. URL https://arxiv.org/abs/1610.05492.

[33] Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-Euclidean trust-region optimization, 2025. URL https://arxiv.org/abs/2503.12645.

[34] Zdislav Kovarik. Some iterative methods for improving orthonormality. *SIAM Journal on Numerical Analysis*, 7(3):386–389, 1970. URL https://doi.org/10.1137/0707031.

[35] Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. *Advances in Neural Information Processing Systems*, 36:40238–40271, 2023. URL https://arxiv.org/abs/2306.01264.

[36] Jiaxiang Li and Mingyi Hong. A note on the convergence of Muon and further, 2025. URL https://arxiv.org/abs/2502.02900.

[37] Feng Liang, Zhen Zhang, Haifeng Lu, Victor Leung, Yanyi Guo, and Xiping Hu. Communication-efficient large-scale distributed deep learning: A comprehensive survey. *arXiv preprint arXiv:2404.06114*, 2024. URL https://arxiv.org/abs/2404.06114.

[38] Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025. URL https://arxiv.org/abs/2502.16982.

[39] Xiaorui Liu, Yao Li, Jiliang Tang, and Ming Yan. A double residual compression algorithm for efficient distributed learning. In *International Conference on Artificial Intelligence and Statistics*, pages 133–143. PMLR, 2020. URL https://arxiv.org/abs/1910.07561.

[40] Yuxing Liu, Rui Pan, and Tong Zhang. AdaGrad under anisotropic smoothness, 2024. URL https://arxiv.org/abs/2406.15244.

[41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

[42] Maksim Makarenko, Elnur Gasanov, Rustem Islamov, Abdurakhmon Sadiev, and Peter Richtárik. Adaptive compression for communication-efficient distributed training. *arXiv preprint arXiv:2211.00188*, 2022. URL https://arxiv.org/abs/2211.00188.

[43] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. URL https://arxiv.org/abs/1602.05629.

[44] Moonshot AI. Kimi K2: Open agentic intelligence, 2025. URL https://moonshotai.github.io/Kimi-K2/.

[45] Arvind Narayanan, Xumiao Zhang, Ruiyang Zhu, Ahmad Hassan, Shuowei Jin, Xiao Zhu, Xiaoxuan Zhang, Denis Rybkin, Zhengxuan Yang, Zhuoqing Morley Mao, et al. A variegated look at 5g in the wild: performance, power, and qoe implications. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, pages 610–625, 2021. URL https://dl.acm.org/doi/10.1145/3452296.3472923.

[46] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012. URL https://epubs.siam.org/doi/10.1137/100802001.

[47] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003. URL https://link.springer.com/book/10.1007/978-1-4419-8853-9.

[48] Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/09c6c3783b4a70054da74f2538ed47c6-Paper.pdf.

[49] Julie Nutini, Issam Laradji, and Mark Schmidt. Let's make Block Coordinate Descent converge faster: Faster greedy rules, message-passing, active-set complexity, and superlinear convergence. *arXiv preprint arXiv:1712.08859*, 2017. URL https://arxiv.org/abs/1712.08859.

[50] Shuo Ouyang, Dezun Dong, Yemao Xu, and Liquan Xiao. Communication optimization strategies for distributed deep neural network training: A survey. *Journal of Parallel and Distributed Computing*, 149:52–65, 2021. ISSN 0743-7315. doi: https://doi.org/10.1016/j.jpdc.2020.11.005. URL https://www.sciencedirect.com/science/article/pii/S0743731520304068.

[51] Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. The FineWeb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024. URL https://arxiv.org/abs/2406.17557.

[52] Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Scion. https://github.com/LIONS-EPFL/scion.git, 2025. GitHub repository.

[53] Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. *arXiv preprint arXiv:2502.07529*, 2025. URL https://arxiv.org/abs/2502.07529.

[54] Thomas Pethick, Wanyun Xie, Mete Erdogan, Kimon Antonakopoulos, Tony Silveti-Falls, and Volkan Cevher. Generalized gradient norm clipping & non-Euclidean (l_0, l_1)-smoothness. *arXiv preprint arXiv:2506.01913*, 2025. URL https://arxiv.org/abs/2506.01913.

[55] Constantin Philippenko and Aymeric Dieuleveut. Preserved central model for faster bidirectional compression in distributed settings. *Advances in Neural Information Processing Systems*, 34:2387–2399, 2021. URL https://arxiv.org/abs/2102.12528.

[56] Vitali Pirau, Aleksandr Beznosikov, Martin Takác, Vladislav Matyukhin, and Alexander V. Gasnikov. Preconditioning meets biased compression for efficient distributed optimization. *Computational Management Science*, 21(1):14, 2024. URL https://doi.org/10.1007/s10287-023-00496-6.

[57] Xun Qian, Peter Richtárik, and Tong Zhang. Error compensated distributed SGD can be accelerated. *Advances in Neural Information Processing Systems*, 34:30401–30413, 2021. URL https://arxiv.org/abs/2010.00091.

[58] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16, 2020. URL https://ieeexplore.ieee.org/document/9355301.

[59] Ahmad Rammal, Kaja Gruntkowska, Nikita Fedin, Eduard Gorbunov, and Peter Richtárik. Communication compression for byzantine robust learning: New efficient algorithms and improved rates. In *International Conference on Artificial Intelligence and Statistics*, pages 1207–1215. PMLR, 2024. URL https://arxiv.org/abs/2310.09804.

[60] Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon. https://github.com/artem-riabinin/

Experiments-estimating-smoothness-for-NanoGPT-and-CNN, 2025. GitHub repository.

[61] Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon: Making Muon & Scion great again! (Bridging theory and practice of LMO-based optimizers for LLMs). *arXiv preprint arXiv:2505.13416*, 2025. URL https://arxiv.org/abs/2505.13416.

[62] Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1): 1–38, 2014. URL https://arxiv.org/abs/1107.2848.

[63] Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. *In Neural Information Processing Systems, 2021.*, 2021. URL https://arxiv.org/abs/2106.05203.

[64] Peter Richtárik, Igor Sokolov, Elnur Gasanov, Ilyas Fatkhullin, Zhize Li, and Eduard Gorbunov. 3PC: Three point compressors for communication-efficient distributed training and a better theory for lazy aggregation. In *International Conference on Machine Learning*, pages 18596–18648. PMLR, 2022. URL https://arxiv.org/abs/2202.00998.

[65] Mher Safaryan, Rustem Islamov, Xun Qian, and Peter Richtárik. FedNL: Making Newton-type methods applicable to federated learning. *arXiv preprint arXiv:2106.02969*, 2021. URL https://arxiv.org/abs/2106.02969.

[66] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014. URL https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/IS140694.pdf.

[67] Ishaan Shah, Anthony M Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J Shah, et al. Practical efficiency of Muon for pretraining. *arXiv preprint arXiv:2505.02222*, 2025. URL https://arxiv.org/abs/2505.02222.

[68] David So, Wojciech Mańke, Hanxiao Liu, Zihang Dai, Noam Shazeer, and Quoc V Le. Searching for efficient transformers for language modeling. *Advances in neural information processing systems*, 34:6010–6022, 2021. URL https://arxiv.org/abs/2109.08668v2.

[69] Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019. URL https://arxiv.org/abs/1909.05350.

[70] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. *Advances in neural information processing systems*, 31, 2018. URL https://arxiv.org/abs/1809.07599.

[71] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. URL https://arxiv.org/abs/2104.09864.

[72] Tao Sun, Qingsong Wang, Dongsheng Li, and Bao Wang. Momentum ensures convergence of SIGNSGD under weaker assumptions. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 33077–33099. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/sun23l.html.

[73] Rafał Szlendak, Alexander Tyurin, and Peter Richtárik. Permutation compressors for provably faster distributed nonconvex optimization. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2110.03300.

[74] Hanlin Tang, Chen Yu, Xiangru Lian, Tong Zhang, and Ji Liu. DoubleSqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. In *International Conference on Machine Learning*, pages 6155–6165. PMLR, 2019. URL https://arxiv.org/abs/1905.05957.

[75] Benjamin Thérien, Xiaolong Huang, Irina Rish, and Eugene Belilovsky. MuLoCo: Muon is a practical inner optimizer for DiLoCo. *arXiv preprint arXiv:2505.23725*, 2025. URL https://arxiv.org/abs/2505.23725.

[76] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. URL https://arxiv.org/abs/2307.09288.

[77] Amund Tveit, Bjørn Remseth, and Arve Skogvold. Muon optimizer accelerates grokking. *arXiv preprint arXiv:2504.16041*, 2025. URL https://arxiv.org/abs/2504.16041.

[78] Alexander Tyurin and Peter Richtárik. DASHA: Distributed nonconvex optimization with communication compression, optimal oracle complexity, and no client synchronization. *11th International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2202.01268.

[79] Alexander Tyurin and Peter Richtárik. 2Direction: Theoretically faster distributed training with bidirectional communication compression. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2305.12379.

[80] Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S Rellermeyer. A survey on distributed machine learning. *Acm computing surveys (csur)*, 53(2):1–33, 2020. URL https://arxiv.org/abs/1912.09789.

[81] Meng Wang, Weijie Fu, Xiangnan He, Shijie Hao, and Xindong Wu. A survey on large-scale machine learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(6):2574–2594, 2020. URL https://arxiv.org/abs/2008.03911.

[82] Yunze Wei, Tianshuo Hu, Cong Liang, and Yong Cui. Communication optimization for distributed training: architecture, advances, and opportunities. *IEEE Network*, 2024. URL https://arxiv.org/abs/2403.07585.

[83] Cong Xie, Shuai Zheng, Sanmi Koyejo, Indranil Gupta, Mu Li, and Haibin Lin. Cser: Communication-efficient SGD with error reset. *Advances in Neural Information Processing Systems*, 33:12593–12603, 2020. URL https://arxiv.org/abs/2007.13221.

[84] Shuo Xie, Mohamad Amin Mohamadi, and Zhiyuan Li. Adam exploits ℓ_∞-geometry of loss landscape via coordinate-wise adaptivity. *arXiv preprint arXiv:2410.08198*, 2024. URL https://arxiv.org/abs/2410.08198.

[85] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017. URL https://arxiv.org/abs/1708.03888.

[86] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019. URL https://arxiv.org/abs/1910.07467.

[87] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1905.11881.

[88] Shuai Zheng, Ziyue Huang, and James Kwok. Communication-efficient distributed blockwise momentum SGD with error-feedback. *Advances in Neural Information Processing Systems*, 32, 2019. URL https://arxiv.org/abs/1905.10936.

Appendix A. Appendix

Contents

Appendix B. Related Work

B.1. Compression

Compression techniques are widely used to reduce the communication and memory costs of optimization algorithms, particularly in large-scale and distributed settings, and have been extensively studied in the Euclidean regime [2, 19, 63, 66]. Rather than transmitting full gradient or parameter vectors, these methods compress them into lower-dimensional or sparse representations, for example through quantization or sparsification [2, 5, 19, 66, 73].

There are two primary compression objectives in distributed optimization: w2s and s2w communication. A large body of prior work focuses exclusively on w2s compression, assuming that broadcasting from the server to the workers is either free or negligible [16, 56, 73, 78]. This assumption is partly due to analytical convenience, but can also be justified in settings where the server has significantly higher bandwidth, greater computational resources, or when the network topology favors fast downlink speeds [25]. However, in many communication environments, this asymmetry does not hold. For instance, in 4G LTE and 5G networks, the upload and download speeds can be comparable, with the ratio between w2s and s2w bandwidths bounded within an order of magnitude [20, 45]. In such cases, s2w communication costs become non-negligible, and optimizing for both directions is essential for practical efficiency [12, 17, 18, 39, 55, 79, 88].

B.2. Error Feedback

To address the communication bottleneck, a natural approach is to apply biased compressors to the transmitted gradients. For the standard (Euclidean) GD, which iterates

$$X^{k+1} = X^k - \gamma^k \nabla f(X^k) = X^k - \gamma^k \left(\frac{1}{n} \sum_{j=1}^{n} \nabla f_j(X^k) \right),$$

where $\gamma^k > 0$ is the stepsize, this would yield the update rule

$$X^{k+1} = X^k - \gamma^k \left(\frac{1}{n} \sum_{j=1}^{n} \mathcal{C}_j^k(\nabla f_j(X^k)) \right).$$

Sadly, this "enhancement" can result in exponential divergence, even in simplest setting of minimizing the average of three strongly convex quadratic functions [5, Example 1]. Empirical evidence of such instability appeared much earlier, prompting Seide et al. [66] to propose a remedy in the form of an *error feedback* (EF) mechanism, which we refer to as EF14.

Initial theoretical insights into EF14 were established in the simpler single-node setting [3, 70]. The method was subsequently analyzed in the convex case by Beznosikov et al. [5], Gorbunov et al. [15], Karimireddy et al. [26]. Next, Qian et al. [57] showed that error feedback methods can be combined with Nesterov-style acceleration [47], though at the cost of incorporating additional unbiased compression, leading to increased communication overhead per iteration. These analyses were later extended to the nonconvex regime by Stich and Karimireddy [69]. This motivated a series of extensions combining error feedback with additional algorithmic components, such as bidirectional compression [74], decentralized training protocols [31], and the incorporation of momentum either on the client [88] or server side [83]. While these works advanced the state of the art, their guarantees relied on strong regularity assumptions, such as bounded gradients (BG) or bounded gradient similarity (BGS), which may be difficult to justify in practical deep learning scenarios.

The limitations of EF14 and its successors were partially overcome by Richtárik et al. [63], who proposed a refined variant termed EF21. EF21 eliminates the need for strong assumptions such as BG and BGS, relying only on standard assumptions (smoothness of the local functions f_j and the existence of a global lower bound on f), while improving the iteration complexity to the desirable $\mathcal{O}(1/\sqrt{K})$ in the deterministic setting. Building on this foundation, a series of extensions and generalizations followed. These include adaptations to partial participation, variance-reduction, proximal setting, and bidirectional compression [12], a generalization from contractive to three-point compressors [64], support for adaptive compression schemes [42], and EF21-P–a modification of EF21 from gradient compression to model compression [17]. Further developments used EF21 in the design of Byzantine robust methods [59], applied it to Hessian communication [22], and extended the theoretical analysis to the (L^0, L^0)–smooth regime [29].

With this historical overview in place, we now narrow our focus to two developments in the error feedback literature that are particularly relevant to this work: EF21 [63] and EF21-P [17].

EF21 is a method for workers-to-server (w2s) (= uplink) communication compression. It aims to solve problem (1) via the iterative process

$$\begin{aligned} X^{k+1} &= X^k - \gamma G^k, \\ G_j^{k+1} &= G_j^k + \mathcal{C}_j^k(\nabla f_j(X^{k+1}) - G_j^k), \\ G^{k+1} &= \frac{1}{n}\sum_{j=1}^n G_j^{k+1}, \end{aligned}$$

where $\gamma > 0$ is the stepsize and $\mathcal{C}_j^k \in \mathbb{B}_2(\alpha_D)$ are independent contractive compressors. In the EF21 algorithm, each client j keeps track of a gradient estimator G_j^k. At each iteration, the clients compute their local gradient $\nabla f_j(X^{k+1})$, subtract the stored estimator G_j^k, and then compresses this difference using a biased compression operator. The compressed update is sent to the server, which aggregates updates from all clients and uses them to update the global model. Concurrently, each client updates its error feedback vector by using the same compressed residual. Importantly, EF21 compresses only the uplink communication (i.e., vectors sent from clients to the server), while downlink communication remains uncompressed. That is, the global model X^{k+1} is transmitted in full precision from the server to all clients, under the assumption that downlink communication is not a bottleneck.

A complementary approach is proposed in the follow-up work of Gruntkowska et al. [17], which introduces a primal variant of EF21, referred to as EF21-P. Unlike EF21, which targets uplink compression (from workers to server), EF21-P is explicitly designed for server-to-workers (s2w) (= downlink) compression. The method proceeds via the iterative scheme

$$\begin{aligned} X^{k+1} &= X^k - \gamma\nabla f(W^k) = X^k - \gamma\frac{1}{n}\sum_{j=1}^n \nabla f_j(W^k), \\ W^{k+1} &= W^k + \mathcal{C}^k(X^{k+1} - W^k), \end{aligned}$$

where $\gamma > 0$ is the stepsize and $\mathcal{C}^k \in \mathbb{B}_2(\alpha_P)$ are independent contractive compressors. Analogous to EF21, the EF21-P method employs error feedback to compensate for the distortion introduced by compression. However, rather than correcting gradient estimates, EF21-P maintains and updates an estimate of the model parameters, W^k. The server computes the update X^{k+1}, but broadcasts only a compressed difference $\mathcal{C}^k(X^{k+1} - W^k)$ to the clients.

In its basic form, EF21-P assumes dense uplink communication–i.e., the clients transmit full gradients $\nabla f_j(W^k)$ to the server. Nonetheless, EF21-P can be naturally extended to bidirectional compression by integrating it with an uplink compression mechanism, enabling full communication efficiency [17].

B.3. Generalized Smoothness

A standard assumption in the convergence analysis of gradient-based methods is Lipschitz smoothness of the gradient (Assumption 3). However, many modern learning problems–especially in deep learning–violate this assumption. Empirical evidence has shown non-smoothness in a variety of architectures and tasks, including `LSTM` language modeling, image classification with `ResNet20` [87], and transformer models [9]. These observations motivated the search for alternative smoothness models that better reflect the behavior of practical objectives.

One such model is (L^0, L^0)–smoothness, introduced by Zhang et al. [87] for twice continuously differentiable functions in the Euclidean setting. The authors define a function $f : \mathbb{R}^d \to \mathbb{R}$ to be (L^0, L^0)–smooth if

$$\left\|\nabla^2 f(X)\right\|_2 \leq L^0 + L^0 \left\|\nabla f(X)\right\|_2 \qquad \forall X \in \mathbb{R}^d.$$

This condition generalizes standard Lipschitz smoothness and has been shown empirically to capture deep learning loss landscapes more faithfully than the classical model [9, 87]. Subsequent works extended the above condition beyond the twice differentiable case [7, 35]. In particular, Chen et al. [7] introduced asymmetric and symmetric variants of (L^0, L^0)–smoothness, where the asymmetric form (a special case of Assumption 4 restricted to Euclidean norms) is given by

$$\left\|\nabla f(X) - \nabla f(Y)\right\|_2 \leq \left(L^0 + L^1 \left\|\nabla f(X)\right\|_2\right) \left\|X - Y\right\|_2 \quad \forall X, Y \in \mathbb{R}^d.$$

This framework has since been used in the non-Euclidean setting [54] and adapted to the layer-wise structure of deep networks by Riabinin et al. [61], who introduced non-Euclidean *layer-wise* (L^0, L^1)–smoothness assumption (Assumption 8). This layer-aware view aligns naturally with LMO-based optimizers that operate on individual parameter groups.

The idea of accounting for the heterogeneous structure of parameters is not unique to the work of Riabinin et al. [61]. Anisotropic smoothness conditions, where smoothness constants can vary across coordinates or parameter blocks, have been studied extensively, for example in the context of coordinate descent methods [46, 49, 62]. Variants of coordinate-wise or block-wise (generalized) smoothness assumptions have also been used to analyze algorithms such as signSGD [4, 9], Adagrad [23, 40], and Adam [84]. These works collectively reinforce the need for smoothness models that reflect the anisotropic geometry of modern neural networks.

Appendix C. Layer-wise Setup

So far, we have been operating in an abstract vector space $\mathcal{S}$, without assuming any particular structure. This is the standard approach in the vast majority of the theoretical optimization literature in ML, where model parameters are typically flattened into vectors in $\mathbb{R}^d$. However, modern deep networks are inherently structured objects, with a clear *layer-wise* organization. While treating parameters as flat vectors can still yield meaningful convergence guarantees, explicitly modeling this layer-wise structure allows us to formulate assumptions that more accurately reflect the underlying geometry of the model Crawshaw et al. [9], Jiang et al. [23], Nesterov [46], Richtárik and Takáč [62]. This, in turn, can lead to improved theoretical results [40, 61].

A further motivation for adopting the layer-wise perspective is that the algorithms that inspired this work–Muon, Scion, and Gluon–are themselves *layer-wise by design*. Rather than operating on the entire parameter vector, they apply separate LMO updates to each layer or building block independently. This modular treatment is one of the main reasons for their strong empirical performance.

With this motivation in mind, we now turn to solving the optimization problem (1) in a setting where the parameter vector $X \in \mathcal{S}$ represents a collection of matrices $X_i \in \mathcal{S}_i := \mathbb{R}^{m_i \times n_i}$ corresponding to the trainable parameters of each layer $i = 1, \dots, p$ in a neural network. For notational convenience, we write $X = [X_1, \dots, X_p]$ and $\nabla f(X) = [\nabla_1 f(X), \dots, \nabla_p f(X)]$, where $\nabla_i f(X)$ is the gradient component corresponding to the ith layer. Accordingly, $\mathcal{S}$ is the d-dimensional product space

$$\mathcal{S} := \bigotimes_{i=1}^{p} \mathcal{S}_i \equiv \mathcal{S}_1 \otimes \cdots \otimes \mathcal{S}_p,$$

where $d := \sum_{i=1}^{p} m_i n_i$. Each component space $\mathcal{S}_i$ is equipped with the trace inner product, defied as $\langle X_i, Y_i \rangle_{(i)} := \operatorname{tr}(X_i^\top Y_i)$ for $X_i, Y_i \in \mathcal{S}_i$, and an arbitrary norm $\|\cdot\|_{(i)}$, not necessarily induced by this inner product. We use $\|\cdot\|_{(i)\star}$ to denote the dual norm associated with $\|\cdot\|_{(i)}$ (i.e., $\|X_i\|_{(i)\star} := \sup_{\|Z_i\|_{(i)} \leq 1} \langle X_i, Z_i \rangle_{(i)}$ for any $X_i \in \mathcal{S}_i$). Furthermore, we use $\underline{\rho}_i, \bar{\rho}_i > 0$ to denote the norm equivalence constants such that

$$\underline{\rho}_i \|X_i\|_{(i)} \leq \|X_i\|_2 \leq \bar{\rho}_i \|X_i\|_{(i)} \qquad \forall X_i \in \mathcal{S}_i,$$

(or, equivalently, $\underline{\rho}_i \|X_i\|_2 \leq \|X_i\|_{(i)\star} \leq \bar{\rho}_i \|X_i\|_2$).

Remark 7 *In the case of* Muon, *the norms $\|\cdot\|_{(i)}$ are taken to be the spectral norms, i.e., $\|\cdot\|_{(i)} = \|\cdot\|_{2\to2}$. Since for any matrix X of rank at most r, we have*

$$\|X\|_{2\to2} \leq \|X\|_F \leq \sqrt{r}\, \|X\|_{2\to2},$$

in this setting, $\underline{\rho}_i = 1$ and $\bar{\rho}_i = \sqrt{r}$.

Given the block structure of X across layers, the smoothness assumptions in Assumption 3 can be made more precise by assigning separate constants to each layer.

Assumption 6 (Layer-wise smoothness) *The function $f : \mathcal{S} \mapsto \mathbb{R}$ is layer-wise L^0–smooth with constants $L^0 := (L_1^0, \dots, L_p^0) \in \mathbb{R}_+^p$, i.e.,*

$$\|\nabla_i f(X) - \nabla_i f(Y)\|_{(i)\star} \leq L_i^0 \|X_i - Y_i\|_{(i)}$$

for all $i = 1, \dots, p$ and all $X = [X_1, \dots, X_p] \in \mathcal{S}$, $Y = [Y_1, \dots, Y_p] \in \mathcal{S}$.

Assumption 7 (Local layer-wise smoothness) *The functions $f_j : \mathcal{S} \mapsto \mathbb{R}$, $j \in [n]$, are layer-wise L_j^0–smooth with constants $L_j^0 := (L_{1,j}^0, \ldots, L_{p,j}^0) \in \mathbb{R}_+^p$, i.e.,*

$$\|\nabla_i f_j(X) - \nabla_i f_j(Y)\|_{(i)\star} \leq L_{i,j}^0 \|X_i - Y_i\|_{(i)}$$

for all $i = 1, \ldots, p$ and all $X = [X_1, \ldots, X_p] \in \mathcal{S}$, $Y = [Y_1, \ldots, Y_p] \in \mathcal{S}$. We define $(\tilde{L}_i^0)^2 := \frac{1}{n}\sum_{j=1}^n (L_{i,j}^0)^2$.

We invoke Assumptions 6 and 7 in Appendices F.3.1 and F.4.1 to extend Theorems 3 and 5 to the more general setting.

Smoothness is the standard assumption used in virtually all convergence results for Muon and Scion [33, 36, 53] (except for the recent work on Gluon [61]). However, as discussed in Section 4 and Section B.3, this assumption often fails to hold in modern deep learning settings. To address this, we adopt a more flexible and expressive condition: the *layer-wise* (L^0, L^1)*–smoothness* assumption [61].

Assumption 8 (Layer-wise (L^0, L^1)–smoothness) *The function $f : \mathcal{S} \mapsto \mathbb{R}$ is layer-wise (L^0, L^1)–smooth with constants $L^0 := (L_1^0, \ldots, L_p^0) \in \mathbb{R}_+^p$ and $L^1 := (L_1^1, \ldots, L_p^1) \in \mathbb{R}_+^p$, i.e.,*

$$\|\nabla_i f(X) - \nabla_i f(Y)\|_{(i)\star} \leq \left(L_i^0 + L_i^1 \|\nabla_i f(X)\|_{(i)\star}\right) \|X_i - Y_i\|_{(i)}$$

for all $i = 1, \ldots, p$ and all $X = [X_1, \ldots, X_p] \in \mathcal{S}$, $Y = [Y_1, \ldots, Y_p] \in \mathcal{S}$.

Since, unlike Gluon, we operate in the distributed setting, we will also need an analogous assumption on the local functions f_j.

Assumption 9 (Local layer-wise (L^0, L^1)–smoothness) *The functions f_j, $j \in [n]$, are layer-wise (L_j^0, L_j^1)–smooth with constants $L_j^0 := (L_{1,j}^0, \ldots, L_{p,j}^0) \in \mathbb{R}_+^p$ and $L_j^1 := (L_{1,j}^1, \ldots, L_{p,j}^1) \in \mathbb{R}_+^p$, i.e.,*

$$\|\nabla_i f_j(X) - \nabla_i f_j(Y)\|_{(i)\star} \leq \left(L_{i,j}^0 + L_{i,j}^1 \|\nabla_i f_j(X)\|_{(i)\star}\right) \|X_i - Y_i\|_{(i)}$$

for all $i = 1, \ldots, p$ and all $X = [X_1, \ldots, X_p] \in \mathcal{S}$, $Y = [Y_1, \ldots, Y_p] \in \mathcal{S}$.

For $O \in \{0, 1\}$, we define $L_{\max,j}^O := \max_{i\in[n]} L_{i,j}^O$, $L_{i,\max}^O := \max_{j\in[n]} L_{i,j}^O$ and $\bar{L}_i^0 := \frac{1}{n}\sum_{j=1}^n L_{i,j}^0$.

Riabinin et al. [61] present empirical evidence showing that this more flexible, layer-wise approach is essential for accurately modeling the network's underlying structure. They demonstrate that the layer-wise (L^0, L^1)–smoothness condition approximately holds along the training trajectory of Gluon in experiments on the `NanoGPT` language modeling task. Motivated by these findings, in Appendices F.3.2 and F.4.2, we provide an analysis within this generalized framework, offering a *full generalization of* Gluon *to bidirectional compression.*

In the stochastic setting, we will also require a layer-wise analogue of Assumption 5.

Assumption 10 *The stochastic gradient estimator $\nabla f(\cdot, \xi) : \mathcal{S} \mapsto \mathcal{S}$ is unbiased and has bounded variance. That is, $\mathbb{E}_{\xi\sim\mathcal{D}}[\nabla f(X, \xi)] = \nabla f(X)$ for all $X \in \mathcal{S}$ and there exists $\sigma_i \geq 0$ such that*

$$\mathbb{E}_{\xi\sim\mathcal{D}}\left[\|\nabla_i f_j(X, \xi) - \nabla_i f_j(X)\|_2^2\right] \leq \sigma_i^2, \quad \forall X \in \mathcal{S},\ i = 1, \ldots, p.$$

We permit layer-dependent variance parameters σ_i^2, motivated by empirical evidence that variance is not uniform across layers. For example, Glentis et al. [14] observe that, during training of `LLaMA 130M` with SGD and column-wise normalization (i.e., Gluon using the $\|\cdot\|_{1\to 2}$ norm), the final and embedding layers display significantly higher variance.

C.1. Muon, Scion and Gluon

Muon, introduced by Jordan et al. [24], is an optimizer for the hidden layers of neural networks (the first and last layers are trained with AdamW). Unlike traditional element-wise gradient methods, it updates each weight matrix as a whole. Given a layer X_i and the corresponding (stochastic) gradient G_i, Muon selects an update that maximizes the alignment with the gradient to reduce loss, while constraining the update's size to avoid excessive model perturbation. This is formulated as a constrained optimization problem over the spectral norm ball:

$$\arg\min_{\Delta X_i} \langle G_i, \Delta X_i \rangle \quad \text{s.t.} \quad \|\Delta X_i\|_{2\to 2} \le t_i, \tag{4}$$

where the radius $t_i > 0$ plays a role similar to a stepsize. The optimal update ΔX_i is obtained by orthogonalizing G_i via its singular value decomposition $G_i = U_i \Sigma_i V_i^T$, leading to

$$\Delta X_i = -t_i U_i V_i^T.$$

This yields the basic update

$$X_i^{k+1} = X_i^k + \Delta X_i^k = X_i^k - t_i^k U_i^k (V_i^k)^T. \tag{5}$$

In practice, computing the SVD exactly at every step is expensive and not GPU-friendly. Muon instead uses Newton–Schulz iterations [6, 34] to approximate the orthogonalization. Combined with momentum, the practical update is

$$M_i^k = (1-\beta_i) M_i^{k-1} + \beta_i G_i^k, \qquad X_i^{k+1} = X_i^k - t_i^k \text{NewtonSchulz}(M_i^k),$$

where $\beta_i \in (0,1]$ is the momentum parameter and M_i^k is the momentum-averaged gradient.

While Newton–Schulz iterations and momentum are crucial for practical efficiency, the essence of Muon lies in solving (4)–that is, computing the linear minimization oracle (LMO) over the spectral norm ball. Recall that $\mathrm{LMO}_{\mathcal{B}(X,t)}(G) := \arg\min_{Z \in \mathcal{B}(X,t)} \langle G, Z \rangle$. Then

$$\Delta X_i = \arg\min_{Z_i \in \mathcal{B}_i^{2\to 2}(0,t_i)} \langle G_i, Z_i \rangle = \mathrm{LMO}_{\mathcal{B}_i^{2\to 2}(0,t_i)}(G_i)$$

where $\mathcal{B}_i^{2\to 2}(0,t_i) := \{Z_i \in \mathcal{S}_i : \|Z_i\|_{2\to 2} \le t_i\}$ is the spectral norm ball of radius t_i around 0. Thus, the update (5) can equivalently be written as

$$X_i^{k+1} = X_i^k + \mathrm{LMO}_{\mathcal{B}_i^{2\to 2}(0,t_i^k)}\left(G_i^k\right), \tag{6}$$

where G_i^k may be replaced with a momentum term.

Crucially, nothing in this formulation ties us to the spectral norm. The same update structure can be defined over any norm ball, opening the door to an entire family of optimizers whose properties depend on the underlying geometry. This insight has led to several Muon-inspired methods

Algorithm 2 Deterministic EF21-Muon

```
Input: radii $t_i^k > 0$ / stepsizes $\gamma_i^k$; initial iterate $X^0 = [X_1^0, \dots, X_p^0] \in \mathcal{S}$ (stored on the server);
       initial iterate shift $W^0 = X^0$ (stored on the server and the workers); initial gradient estimators
       $G_j^0 = [G_{1,j}^0, \dots, G_{p,j}^0] = [\nabla_1 f_j(X^0), \dots, \nabla_p f_j(X^0)] \in \mathcal{S}$ (stored on the workers),
       $G^0 = \frac{1}{n}\sum_{j=1}^n G_j^0$ (stored on the server); worker compressors $\mathcal{C}_i^k$; server compressors $\mathcal{C}^k$
for $k = 0, 1, \dots, K-1$ do
    for $i = 1, 2, \dots, p$ do
        $X_i^{k+1} = \mathrm{LMO}_{\mathcal{B}(X_i^k, t_i^k)}\left(G_i^k\right) = X_i^k - \gamma_i^k (G_i^k)^\sharp$    // Take LMO-type step
        $S_i^k = \mathcal{C}_i^k(X_i^{k+1} - W_i^k)$    // Compress shifted model on the server
        $W_i^{k+1} = W_i^k + S_i^k$    // Update model shift
        Broadcast $S^k = [S_1^k, \dots, S_p^k]$ to all workers
    end
    for $j = 1, \dots, n$ in parallel do
        for $i = 1, 2, \dots, p$ do
            $W_i^{k+1} = W_i^k + S_i^k$    // Update model shift
            $R_{i,j}^{k+1} = \mathcal{C}_{i,j}^k(\nabla_i f_j(W^{k+1}) - G_{i,j}^k)$    // Compress shifted gradient
            $G_{i,j}^{k+1} = G_{i,j}^k + R_{i,j}^{k+1}$
        end
        Broadcast $R_j^{k+1} = [R_{1,j}^{k+1}, \dots, R_{p,j}^{k+1}]$ to the server
    end
    for $i = 1, \dots, p$ do
        $G_i^{k+1} = \frac{1}{n}\sum_{j=1}^n G_{i,j}^{k+1} = G_i^k + \frac{1}{n}\sum_{j=1}^n R_{i,j}^{k+1}$    // Compute gradient estimator
    end
end
```

with provable convergence guarantees [33, 53, 61]. Scion [53] removes the restriction to matrix-shaped layers by applying LMO-based updates to all layers, pairing the spectral norm for hidden layers with the $\|\cdot\|_{1\to\infty}$ norm elsewhere. Gluon [61] expands the view even further: it provides a general convergence analysis for LMO updates over arbitrary norm balls, supported by a layer-wise (L^0, L^1)–-smoothness assumption that captures the heterogeneity of deep learning loss landscapes more accurately than standard smoothness.

C.2. Layer-wise EF21-Muon

The simplified EF21-Muon algorithm in Algorithm 1, analyzed in Section 4 omits the structured, layer-wise treatment introduced above. richer setting is presented in Algorithm 3. Moreover, in Algorithm 2, we formalize the deterministic counterpart of EF21-Muon, whose simplified variant we analyzed in Section 4.

Both Algorithms 2 and 3 operate on a per-layer basis. We now briefly describe their structure. For each layer i, the parameters are updated via $X_i^{k+1} = \mathrm{LMO}_{\mathcal{B}(X_i^k, t_i^k)}\left(G_i^k\right)$ (equivalently, $X_i^{k+1} = X_i^k - \gamma_i^k \left(G_i^k\right)^\sharp$, where $\gamma^k = {}^{t^k}/_{\|G^k\|_\star}$–see Section D). Next, the algorithms perform the server-to-workers (s2w) compression, following a technique inspired by EF21-P [17]. The resulting

Algorithm 3 EF21-Muon

```
Input: radii $t_i^k > 0$ / stepsizes $\gamma_i^k$; momentum parameters $\beta_i \in (0,1]$; initial iterate $X^0 = [X_1^0, \dots, X_p^0] \in \mathcal{S}$ (stored on the server); initial iterate shift $W^0 = X^0$ (stored on the server and the workers); initial gradient estimators $G_j^0 = [G_{1,j}^0, \dots, G_{p,j}^0] \in \mathcal{S}$ (stored on the workers); $G^0 = \frac{1}{n}\sum_{j=1}^n G_j^0$ (stored on the server); initial momentum $M_j^0 = [M_{1,j}^0, \dots, M_{p,j}^0] \in \mathcal{S}$ (stored on the workers); worker compressors $\mathcal{C}_{i,j}^k$; server compressors $\mathcal{C}_i^k$
for $k = 0, 1, \dots, K-1$ do
    for $i = 1, 2, \dots, p$ do
        $X_i^{k+1} = \mathrm{LMO}_{\mathcal{B}(X_i^k, t_i^k)}\left(G_i^k\right) = X_i^k - \gamma_i^k (G_i^k)^\sharp$    // Take LMO-type step
        $S_i^k = \mathcal{C}_i^k(X_i^{k+1} - W_i^k)$    // Compress shifted model on the server
        $W_i^{k+1} = W_i^k + S_i^k$    // Update model shift
        Broadcast $S^k = [S_1^k, \dots, S_p^k]$ to all workers
    end
    for $j = 1, \dots, n$ in parallel do
        for $i = 1, 2, \dots, p$ do
            $W_i^{k+1} = W_i^k + S_i^k$    // Update model shift
            $M_{i,j}^{k+1} = (1-\beta_i)M_{i,j}^k + \beta_i \nabla_i f_j(W^{k+1}, \xi_j^{k+1})$    // Compute momentum
            $R_{i,j}^{k+1} = \mathcal{C}_{i,j}^k(M_{i,j}^{k+1} - G_{i,j}^k)$    // Compress shifted gradient
            $G_{i,j}^{k+1} = G_{i,j}^k + R_{i,j}^{k+1}$
        end
        Broadcast $R_j^{k+1} = [R_{1,j}^{k+1}, \dots, R_{p,j}^{k+1}]$ to the server
    end
    for $i = 1, 2, \dots, p$ do
        $G_i^{k+1} = \frac{1}{n}\sum_{j=1}^n G_{i,j}^{k+1} = G_i^k + \frac{1}{n}\sum_{j=1}^n R_{i,j}^{k+1}$    // Compute gradient estimator
    end
end
```

compressed messages $S_i^k = \mathcal{C}_i^k(X_i^{k+1} - W_i^k)$ are sent to the workers. Each worker then updates the model shift and uses the resulting model estimate W_i^{k+1} to compute the local (stochastic) gradient. This gradient is then used (either directly or within a momentum term) to form the compressed message $R_{i,j}^{k+1}$. This part of the algorithm follows the workers-to-server (w2s) compression strategy of EF21 [63]. The messages $R_{i,j}^{k+1}$ are sent back to the server, which updates the layer-wise gradient estimators via $G_i^{k+1} = \frac{1}{n}\sum_{j=1}^n G_{i,j}^{k+1} = G_i^k + \frac{1}{n}\sum_{j=1}^n R_{i,j}^{k+1}$. This process is repeated until convergence.

Appendix D. LMO in Many Guises

As outlined in Section 2, the update rule (2)

$$X^{k+1} = X^k + t^k \mathrm{LMO}_{\mathcal{B}(0,1)}\left(G^k\right)$$

admits several equivalent reformulations.

LMO viewpoint. The original update (2) is the solution of a simple linear minimization problem over a norm ball

$$X^{k+1} = \mathrm{LMO}_{\mathcal{B}(X^k,t^k)}\left(G^k\right) = \underset{X\in\mathcal{B}(X^k,t^k)}{\arg\min} \left\langle G^k, X\right\rangle,$$

where $\mathcal{B}(X,t) := \{Z \in \mathcal{S} \ : \ \|Z - X\| \leq t\}$. The LMO satisfies

$$\left\langle G, \mathrm{LMO}_{\mathcal{B}(X,t)}\left(G\right)\right\rangle = -t\left\|G\right\|_\star.$$

Sharp operator viewpoint. An equivalent perspective is obtained via the *sharp operators* [28, 46]. Define the function $\phi(X) := \frac{1}{2}\left\|X\right\|^2$. Its Fenchel conjugate is given by

$$\phi^\star(G) := \sup_{X\in\mathcal{S}} \left\{\langle G, X\rangle - \phi(X)\right\} = \frac{1}{2}\left\|X\right\|_\star^2,$$

and its subdifferential $\partial\phi^\star$ coincides with the sharp operator:

$$\begin{aligned}\partial\phi^\star(G) &= \left\{X \in \mathcal{S} : \langle G, X\rangle = \left\|G\right\|_\star\left\|X\right\|, \left\|G\right\|_\star = \left\|X\right\|\right\}\\ &= -\left\|G\right\|_\star \mathrm{LMO}_{\mathcal{B}(0,1)}\left(G\right)\\ &= G^\sharp,\end{aligned}$$

where $G^\sharp := \arg\max_{X\in\mathcal{S}}\{\langle G, X\rangle - \frac{1}{2}\left\|X\right\|^2\}$ is the *sharp operator*. Therefore,

$$X^{k+1} = X^k + t^k \mathrm{LMO}_{\mathcal{B}(0,1)}\left(G^k\right) = X^k - \frac{t^k}{\left\|G^k\right\|_\star}\left(G^k\right)^\sharp,$$

i.e., a normalized steepest descent step with effective stepsize $\gamma^k := t^k/\left\|G^k\right\|_\star$.

Two properties of the sharp operator used later are

$$\left\langle X, X^\sharp\right\rangle = \left\|X^\sharp\right\|^2, \qquad \left\|X\right\|_\star = \left\|X^\sharp\right\|.$$

Subdifferential viewpoint. The negative LMO direction $-\mathrm{LMO}_{B(0,1)}\left(A\right) = \arg\max_{\|Z\|=1}\langle A, Z\rangle$ is a subdifferential of the dual norm $\partial\left\|\cdot\right\|_\star(A)$, so (2) can also be written as

$$X^{k+1} = X^k + t^k \mathrm{LMO}_{B(0,1)}\left(G^k\right) = X^k - t^k H^k$$

for some $H^k \in \partial\left\|\cdot\right\|_\star(G^k)$, where by the definition of subdifferential, for any $G^k \neq 0$,

$$\left\langle H^k, G^k\right\rangle = \left\|G^k\right\|_\star, \quad \left\|H^k\right\| = 1. \tag{7}$$

Appendix E. Non-Euclidean Contractive Compressors

Recall from Theorem 1 that a mapping $\mathcal{C} : \mathcal{S} \to \mathcal{S}$ is called a *contractive compression operator* with parameter $\alpha \in (0, 1]$ if, for all $X \in \mathcal{S}$,

$$\mathbb{E}\left[\|\mathcal{C}(X) - X\|^2\right] \leq (1 - \alpha)\|X\|^2 . \tag{8}$$

When $\|\cdot\|$ is the Euclidean norm, a wide range of such compressors is available in the literature [2, 5, 19, 63, 66, 73]. However, when $\|\cdot\|$ is a *non-Euclidean* norm, Euclidean contractivity does not in general imply contractivity with respect to $\|\cdot\|$. Indeed, suppose that $\mathcal{C}$ is contractive with respect to the Euclidean norm. Then, using norm equivalence, for any $X \in \mathcal{S}$,

$$\underline{\rho}^2\mathbb{E}\left[\|\mathcal{C}(X) - X\|^2\right] \leq \mathbb{E}\left[\|\mathcal{C}(X) - X\|_2^2\right] \leq (1 - \alpha)\|X\|_2^2 \leq \bar{\rho}^2(1 - \alpha)\|X\|^2 .$$

Rearranging gives

$$\mathbb{E}\left[\|\mathcal{C}(X) - X\|^2\right] \leq \frac{\bar{\rho}^2}{\underline{\rho}^2}(1 - \alpha)\|X\|^2 ,$$

and hence $\mathcal{C}$ is *not* contractive with respect to the norm $\|\cdot\|$ unless $\alpha > 1 - \underline{\rho}^2/\bar{\rho}^2$. Consequently, dedicated compressors are needed when working outside the Euclidean setting.

In this section, we first present two simple examples of operators that satisfy condition (8) for *any* norm. These are, however, in general not very practical choices. We then turn to more useful examples of non-Euclidean compressors for several matrix norms of interest.

A simple deterministic example of a contractive compressor is the *scaling* or *damping* operator.

Definition 8 (Deterministic Damping) *For any $X \in \mathcal{S}$, the deterministic damping operator with parameter $\gamma \in (0, 2)$ is defined as*

$$\mathcal{C}(X) = \gamma X.$$

For this operator,

$$\mathbb{E}\left[\|\mathcal{C}(X) - X\|^2\right] = (1 - \gamma)^2\|X\|^2 ,$$

and thus $\mathcal{C}$ satisfies Theorem 1 with $\alpha = 1 - (1 - \gamma)^2$ for any $\gamma \in (0, 2)$.

Despite meeting the definition, the deterministic damping operator is of little use in communication-constrained optimization: it merely scales the entire input vector by a constant, without reducing the amount of data to be transmitted. The fact that it formally satisfies the contractive compressor definition is more of a theoretical curiosity. It highlights that the definition captures a broader mathematical property that does not always align with the practical engineering goal of reducing data transmission.

The *random dropout operator* (whose scaled, unbiased variant appears in the literature as the *Bernoulli compressor* [21]) is a simple yet more practically relevant example of a contractive compressor that can reduce communication cost.

Definition 9 (Random Dropout) *For any $X \in \mathcal{S}$, the random dropout operator with a probability parameter $p \in (0,1]$ is defined as*

$$\mathcal{C}(X) = \begin{cases} X & \textit{with probability } p, \\ 0 & \textit{with probability } 1-p. \end{cases}$$

Then

$$\mathbb{E}\left[\|\mathcal{C}(X) - X\|^2\right] = (1-p)\|X\|^2,$$

and hence $\mathcal{C} \in \mathbb{B}(p)$.

The examples of deterministic damping and random dropout apply to any valid norm defined on the space $\mathcal{S}$. However, one can also design compressors directly for the norm of interest. A natural example for both the spectral norm $\|\cdot\|_{2\to2}$ and the nuclear norm $\|\cdot\|_*$ is based on truncated SVD.

Definition 10 (Top$K$ SVD compressor) *Let $X = U\Sigma V^\top \in \mathbb{R}^{m\times n}$ be a matrix of rank r, where $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ contains the singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$. For $K < r$, the* Top$K$ SVD compressor *is defined by*

$$\mathcal{C}(X) := U\Sigma_K V^\top,$$

where where $\Sigma_K = \operatorname{diag}(\sigma_1, \ldots, \sigma_K, 0, \ldots, 0)$ retains the K largest singular values, setting the rest to zero.

The Top$K$ SVD compressor can be used in conjunction with several commonly used matrix norms:

- **Spectral norm.** The spectral norm, frequently used in LMO-based optimization methods, is defined by $\|X\|_{2\to2} = \sigma_1$. Under this norm, the compression residual is

  $$\|X - \mathcal{C}(X)\|_{2\to2} = \sigma_{K+1}.$$

  This yields a valid contractive compressor (unless $\sigma_{K+1}^2 = \sigma_1^2$), and Theorem 1 is satisfied with parameter $\alpha = 1 - \sigma_{K+1}^2/\sigma_1^2$.

- **Nuclear norm.** The nuclear norm, dual to the spectral norm, is given by $\|X\|_* = \sum_{i=1}^r \sigma_i$. In this case,

  $$\|X - \mathcal{C}(X)\|_* = \sum_{i=K+1}^r \sigma_i,$$

  and Theorem 1 holds with $\alpha = 1 - \left(\frac{\sum_{i=K+1}^r \sigma_i}{\sum_{i=1}^r \sigma_i}\right)^2$.

- **Frobenius norm.** The Euclidean norm of the matrix can be expressed as $\|X\|_F = \sqrt{\sum_{i=1}^r \sigma_i^2}$. Then,

  $$\|X - \mathcal{C}(X)\|_F = \sqrt{\sum_{i=K+1}^r \sigma_i^2}.$$

  and so Theorem 1 is satisfied with $\alpha = 1 - \frac{\sum_{i=K+1}^r \sigma_i^2}{\sum_{i=1}^r \sigma_i^2}$.

In fact, the Top$K$ SVD compressor is naturally well-suited for a larger family of *Schatten p-norm*, defined in terms of the singular values σ_i of a matrix X by

$$\|X\|_{S_p} = \left(\sum_{i=1}^{r} \sigma_i^p\right)^{1/p}$$

Important special cases include the nuclear norm (or trace norm) for $p = 1$ (i.e., $\|X\|_* = \|X\|_{S_1}$), the Frobenius norm for $p = 2$ (i.e., $\|X\|_F = \|X\|_{S_2}$), and the spectral norm for $p = \infty$ (i.e., $\|X\|_{2\to 2} = \|X\|_{S_\infty}$). In general, it is easy to show that the Top$K$ SVD compressor satisfies Theorem 1 with respect to the $\|\cdot\|_{S_p}$ norm with

$$\alpha = 1 - \left(\frac{\sum_{i=K+1}^{r} \sigma_i^p}{\sum_{i=1}^{r} \sigma_i^p}\right)^{2/p}.$$

Remark 11 *For large-scale matrices, computing the exact SVD may be computationally prohibitive. In such cases, one may resort to approximate methods to obtain a stochastic compressor $\widetilde{\mathcal{C}}$ satisfying Theorem 1* in expectation*:*

$$\mathbb{E}\left[\left\|\widetilde{\mathcal{C}}(X) - X\right\|^2\right] \leq (1 - \alpha + \delta)\ \|X\|^2,$$

where $\delta > 0$ quantifies the approximation error and can be made arbitrarily small.

Remark 12 *The expressions for α above depend on the singular values of X, and hence α is generally* matrix-dependent *rather than a uniform constant. For theoretical guarantees, one may take the minimum α observed over a training run. Alternatively, our framework admits a straightforward extension to iteration-dependent compression parameters.*

Beyond Schatten norms, similar ideas can be applied to other structured non-Euclidean norms. Throughout, we let $X_{i:}$, $X_{:j}$, and X_{ij} denote the ith row, jth column, and (i, j)th entry of the matrix $X \in \mathbb{R}^{m\times n}$, respectively.

Definition 13 (Column-wise Top$_p K$ compressor) *The* column-wise Top$_p K$ compressor *keeps the K columns with largest ℓ_p norm, setting the rest to zero:*

$$\mathcal{C}(X)_{:j} = \begin{cases} X_{:j}, & j \in \mathcal{I}_K, \\ 0, & \textit{otherwise}, \end{cases}$$

where $\mathcal{I}_K$ indexes the K columns with the largest ℓ_p norm.

This operator is naturally suited for the mixed $\ell_{p,q}$ norms ($p, q \geq 1$), defined as

$$\|X\|_{p,q} := \left(\sum_{j=1}^{n}\left(\sum_{i=1}^{m} |X_{ij}|^p\right)^{q/p}\right)^{1/q} = \left(\sum_{j=1}^{n} \|X_{:j}\|_p^q\right)^{1/q},$$

where $\|\cdot\|_p$ is the standard (vector) ℓ_p norm. The compression residual satisfies

$$\|X - \mathcal{C}(X)\|_{p,q} = \left(\sum_{j \notin \mathcal{I}_K} \|X_{:j}\|_p^q \right)^{1/q},$$

and hence Theorem 1 holds with

$$\alpha = 1 - \left(\frac{\sum_{j \notin \mathcal{I}_K} \|X_{:j}\|_p^q}{\sum_{j=1}^n \|X_{:j}\|_p^q} \right)^{2/q}.$$

This general formulation recovers, for example, the $\ell_{2,1}$ norm (commonly used in robust data analysis [48]) and the $\ell_{2,2}$ norm (Frobenius norm).

E.1. Compression via Norm Selection

A useful perspective on communication reduction in distributed optimization emerges from the connection between compression operators and mappings such as the *sharp operator* and the LMO. Recall that for any norm $\|\cdot\|$ on a vector space $\mathcal{S}$ with dual norm $\|\cdot\|_*$, the sharp operator of $G \in \mathcal{S}$ is defined as

$$G^\sharp := \arg\max_{X \in \mathcal{S}} \left\{ \langle G, X \rangle - \frac{1}{2} \|X\|^2 \right\}.$$

Since $\|G\|_\star \, \mathrm{LMO}_{\mathcal{B}(0,1)}(G) = -G^\sharp$, one can view $G^\sharp$ as the LMO over the unit ball of $\|\cdot\|$, scaled by $\|G\|_\star$.

For many norms, $G^\sharp$ naturally acts as a structured compressor. Below, we list several such examples.

- **Nuclear norm.** For the nuclear norm $\|X\|_* = \sum_{j=1}^r \sigma_j$ (with dual norm $\|\cdot\|_{2\to2}$, the operator/spectral norm), the sharp operator is

  $$G^\sharp = \sigma_1 \, u_1 v_1^\top,$$

  where σ_1, u_1, and v_1 are the leading singular value and singular vectors of G, yielding a *Rank*1 *compression* via truncated SVD. This operator satisfies Theorem 1 with parameter $\alpha = 1/r$, where r is the rank of G.

- **Element-wise ℓ_1 norm.** For the norm $\|X\|_1 = \sum_{i=1}^m \sum_{j=1}^n |X_{ij}|$ (with dual $\|X\|_\infty = \max_{i,j} |X_{ij}|$), the sharp operator is

  $$G^\sharp = \mathrm{Top1}(G) = \|G\|_\infty \, E_{(i^\star j^\star)},$$

  where $(i^\star, j^\star) = \arg\max_{i,j} |G_{ij}|$ and $E_{(i^\star j^\star)}$ is the matrix with a 1 in entry $(i^\star, j^\star)$ and zeros elsewhere. Thus, the sharp operator associated with the ℓ_1 norm corresponds to *Top*1 *sparsification*, which satisfies Theorem 1 with $\alpha = 1/mn$.

- **Max row sum norm.** For $\|X\|_{\infty\to\infty} = \max_{1\leq i\leq m}\sum_{j=1}^{n}|X_{ij}|$, the dual norm is $\|X\|_{1,\infty} = \sum_{j=1}^{n}\|X_{:j}\|_{\infty}$, and the sharp operator yields

$$G^{\sharp} = \left(\sum_{j=1}^{n}\|X_{:j}\|_{\infty}\right)[\mathrm{sign}(\mathrm{Top1}(G_{:1})),\ldots,\mathrm{Top1}(G_{:b}))]$$

  i.e., it keeps a single non-zero entry in each column of G, with all of these entries equal across columns.

These are only some examples of the compression capabilities of sharp operators. They open the door to compressed server-to-worker communication even in the absence of primal compression, as briefly mentioned in Section 4. Indeed, instead of broadcasting the compressed messages S^k in Algorithms 1 to 3, the server can compute $G^{\sharp}$, transmit this naturally compressed object, and let the workers perform the model update locally. In doing so, we preserve communication efficiency while avoiding the introduction of additional primal compressors.

Appendix F. Convergence Analysis

F.1. Descent Lemmas

We provide two descent lemmas corresponding to the two smoothness regimes. The first applies to the layer-wise smooth setting.

Lemma 14 (Descent Lemma I) *Let Assumption 6 hold and consider the update rule* $X_i^{k+1} = X_i^k - \gamma_i^k \left(G_i^k\right)^\sharp$, $i = 1, \dots, p$, *where* $X^{k+1} = [X_1^{k+1}, \dots, X_p^{k+1}], X^k = [X_1^k, \dots, X_p^k], G^k = [G_1^k, \dots, G_p^k] \in \mathcal{S}$ *and* $\gamma_i^k > 0$. *Then*

$$\begin{aligned} f(X^{k+1}) &\leq f(X^k) + \sum_{i=1}^p \frac{3\gamma_i^k}{2} \left\| \nabla_i f(X^k) - G_i^k \right\|_{(i)\star}^2 - \sum_{i=1}^p \frac{\gamma_i^k}{4} \left\| \nabla_i f(X^k) \right\|_{(i)\star}^2 \\ &\quad - \sum_{i=1}^p \left(\frac{1}{4\gamma_i^k} - \frac{L_i^0}{2} \right) (\gamma_i^k)^2 \left\| G_i^k \right\|_{(i)\star}^2 . \end{aligned}$$

Proof First, for any $s > 0$, we have

$$\begin{aligned} \left\| \nabla f(X^k) \right\|_{(i)\star}^2 &= \left\| \nabla f(X^k) - G^k + G^k \right\|_{(i)\star}^2 \\ &\overset{(27)}{\leq} (1+s) \left\| \nabla f(X^k) - G^k \right\|_{(i)\star}^2 + \left(1 + \frac{1}{s}\right) \left\| G^k \right\|_{(i)\star}^2 , \end{aligned}$$

meaning that

$$\begin{aligned} -\left\| G^k \right\|_{(i)\star}^2 &\leq \frac{1+s}{1+\frac{1}{s}} \left\| \nabla f(X^k) - G^k \right\|_{(i)\star}^2 - \frac{1}{1+\frac{1}{s}} \left\| \nabla f(X^k) \right\|_{(i)\star}^2 \\ &= s \left\| \nabla f(X^k) - G^k \right\|_{(i)\star}^2 - \frac{s}{s+1} \left\| \nabla f(X^k) \right\|_{(i)\star}^2 . \end{aligned} \tag{9}$$

Then, using layer-wise smoothness of f and Theorem 41 with $L_i^1 = 0$, we get

$$\begin{aligned} f(X^{k+1}) &\leq f(X^k) + \left\langle \nabla f(X^k), X^{k+1} - X^k \right\rangle + \sum_{i=1}^p \frac{L_i^0}{2} \left\| X_i^k - X_i^{k+1} \right\|_{(i)}^2 \\ &= f(X^k) + \sum_{i=1}^p \left\langle \nabla_i f(X^k), X_i^{k+1} - X_i^k \right\rangle_{(i)} + \sum_{i=1}^p \frac{L_i^0}{2} \left\| X_i^k - X_i^{k+1} \right\|_{(i)}^2 \\ &= f(X^k) - \sum_{i=1}^p \gamma_i^k \left\langle \nabla_i f(X^k) - G_i^k, \left(G_i^k\right)^\sharp \right\rangle_{(i)} - \sum_{i=1}^p \gamma_i^k \left\langle G_i^k, \left(G_i^k\right)^\sharp \right\rangle_{(i)} \\ &\quad + \sum_{i=1}^p \frac{L_i^0}{2} (\gamma_i^k)^2 \left\| \left(G_i^k\right)^\sharp \right\|_{(i)\star}^2 \\ &\overset{(32),(33)}{=} f(X^k) - \sum_{i=1}^p \gamma_i^k \left\langle \nabla_i f(X^k) - G_i^k, \left(G_i^k\right)^\sharp \right\rangle_{(i)} - \sum_{i=1}^p \frac{\gamma_i^k}{2} \left\| G_i^k \right\|_{(i)\star}^2 \\ &\quad - \sum_{i=1}^p \frac{\gamma_i^k}{2} \left\| G_i^k \right\|_{(i)\star}^2 + \sum_{i=1}^p \frac{L_i^0}{2} (\gamma_i^k)^2 \left\| G_i^k \right\|_{(i)\star}^2 \end{aligned}$$

$$
\begin{aligned}
&\overset{(9)}{\leq} \quad f(X^k) - \sum_{i=1}^p \gamma_i^k \left\langle \nabla_i f(X^k) - G_i^k, \left(G_i^k\right)^\sharp \right\rangle - \sum_{i=1}^p \frac{\gamma_i^k}{2} \left\| G_i^k \right\|_{(i)\star}^2 \\
&\qquad + \sum_{i=1}^p \frac{\gamma_i^k}{2} s \left\| \nabla_i f(X^k) - G_i^k \right\|_{(i)\star}^2 - \sum_{i=1}^p \frac{\gamma_i^k}{2} \frac{s}{s+1} \left\| \nabla_i f(X^k) \right\|_{(i)\star}^2 \\
&\qquad + \sum_{i=1}^p \frac{L_i^0}{2} (\gamma_i^k)^2 \left\| G_i^k \right\|_{(i)\star}^2 .
\end{aligned}
$$

Therefore, applying Fenchel's inequality, we get

$$
\begin{aligned}
& f(X^{k+1}) \\
&\overset{(28)}{\leq} f(X^k) + \sum_{i=1}^p \left(\frac{\gamma_i^k}{2r} \left\| \nabla_i f(X^k) - G_i^k \right\|_{(i)\star}^2 + \frac{\gamma_i^k r}{2} \left\| \left(G_i^k\right)^\sharp \right\|_{(i)}^2 - \frac{\gamma_i^k}{2} \left\| G_i^k \right\|_{(i)\star}^2 \right. \\
&\qquad \left. + \frac{\gamma_i^k}{2} s \left\| \nabla_i f(X^k) - G_i^k \right\|_{(i)\star}^2 - \frac{\gamma_i^k}{2} \frac{s}{s+1} \left\| \nabla_i f(X^k) \right\|_{(i)\star}^2 + \frac{L_i^0}{2} (\gamma_i^k)^2 \left\| G_i^k \right\|_{(i)\star}^2 \right) \\
&\overset{(33)}{=} f(X^k) + \sum_{i=1}^p \left(\frac{\gamma_i^k}{2r} + \frac{\gamma_i^k s}{2} \right) \left\| \nabla_i f(X^k) - G_i^k \right\|_{(i)\star}^2 - \sum_{i=1}^p \frac{\gamma_i^k}{2} \frac{s}{s+1} \left\| \nabla_i f(X^k) \right\|_{(i)\star}^2 \\
&\qquad - \sum_{i=1}^p \left(\frac{1-r}{2\gamma_i^k} - \frac{L_i^0}{2} \right) (\gamma_i^k)^2 \left\| G_i^k \right\|_{(i)\star}^2
\end{aligned}
$$

for any $r > 0$. Choosing $s = 1$ and $r = 1/2$ finishes the proof. ■

The next lemma is specific to the layer-wise smooth case.

Lemma 15 (Descent Lemma II) *Let Assumption 8 hold and consider the update rule* $X_i^{k+1} = \mathrm{LMO}_{\mathcal{B}(X_i^k, t_i^k)}\left(G_i^k\right)$, $i = 1, \ldots, p$, *where* $X^{k+1} = [X_1^{k+1}, \ldots, X_p^{k+1}]$, $X^k = [X_1^k, \ldots, X_p^k]$, $G^k = [G_1^k, \ldots, G_p^k] \in \mathcal{S}$ *and* $t_i^k > 0$. *Then*

$$
\begin{aligned}
f(X^{k+1}) \quad \leq \quad & f(X^k) + \sum_{i=1}^p 2t_i^k \left\| \nabla_i f(X^k) - G_i^k \right\|_{(i)\star} - \sum_{i=1}^p t_i^k \left\| \nabla_i f(X^k) \right\|_{(i)\star} \\
& + \sum_{i=1}^p \frac{L_i^0 + L_i^1 \left\| \nabla_i f(X^k) \right\|_{(i)\star}}{2} (t_i^k)^2 .
\end{aligned}
$$

Proof Assumption 8 and Theorem 41 give

$$
\begin{aligned}
& f(X^{k+1}) \\
& \leq \quad f(X^k) + \left\langle \nabla f(X^k), X^{k+1} - X^k \right\rangle + \sum_{i=1}^p \frac{L_i^0 + L_i^1 \left\| \nabla_i f(X^k) \right\|_{(i)\star}}{2} \left\| X_i^k - X_i^{k+1} \right\|_{(i)}^2 \\
& = \quad f(X^k) + \sum_{i=1}^p \left\langle \nabla_i f(X^k), X_i^{k+1} - X_i^k \right\rangle_{(i)} + \sum_{i=1}^p \frac{L_i^0 + L_i^1 \left\| \nabla_i f(X^k) \right\|_{(i)\star}}{2} \left\| X_i^k - X_i^{k+1} \right\|_{(i)}^2
\end{aligned}
$$

$$
\begin{aligned}
&= f(X^k) + \sum_{i=1}^{p} \left(\left\langle \nabla_i f(X^k) - G_i^k, X_i^{k+1} - X_i^k \right\rangle_{(i)} + \left\langle G_i^k, X_i^{k+1} - X_i^k \right\rangle_{(i)} \right) \\
&\quad + \sum_{i=1}^{p} \frac{L_i^0 + L_i^1 \left\| \nabla_i f(X^k) \right\|_{(i)\star}}{2} (t_i^k)^2 \\
&\overset{(31)}{=} f(X^k) + \sum_{i=1}^{p} \left(\left\langle \nabla_i f(X^k) - G_i^k, X_i^{k+1} - X_i^k \right\rangle_{(i)} - t_i^k \left\| G_i^k \right\|_{(i)\star} \right) \\
&\quad + \sum_{i=1}^{p} \frac{L_i^0 + L_i^1 \left\| \nabla_i f(X^k) \right\|_{(i)\star}}{2} (t_i^k)^2 \\
&\leq f(X^k) + \sum_{i=1}^{p} \left(t_i^k \left\| \nabla_i f(X^k) - G_i^k \right\|_{(i)\star} - t_i^k \left\| G_i^k \right\|_{(i)\star} + \frac{L_i^0 + L_i^1 \left\| \nabla_i f(X^k) \right\|_{(i)\star}}{2} (t_i^k)^2 \right),
\end{aligned}
$$

where the last line follows from the Cauchy-Schwarz inequality and the fact that $\left\| X_i^{k+1} - X_i^k \right\|_{(i)} = t_i^k$. Therefore, using triangle inequality, we get

$$
\begin{aligned}
&f(X^{k+1}) \\
&\leq f(X^k) + \sum_{i=1}^{p} \left(t_i^k \left\| \nabla_i f(X^k) - G_i^k \right\|_{(i)\star} + t_i^k \left\| \nabla_i f(X^k) - G_i^k \right\|_{(i)\star} - t_i^k \left\| \nabla_i f(X^k) \right\|_{(i)\star} \right) \\
&\quad + \sum_{i=1}^{p} \frac{L_i^0 + L_i^1 \left\| \nabla_i f(X^k) \right\|_{(i)\star}}{2} (t_i^k)^2 \\
&= f(X^k) + \sum_{i=1}^{p} \left(2 t_i^k \left\| \nabla_i f(X^k) - G_i^k \right\|_{(i)\star} - t_i^k \left\| \nabla_i f(X^k) \right\|_{(i)\star} \right) \\
&\quad + \sum_{i=1}^{p} \frac{L_i^0 + L_i^1 \left\| \nabla_i f(X^k) \right\|_{(i)\star}}{2} (t_i^k)^2.
\end{aligned}
$$

■

F.2. Auxiliary Lemmas

Lemma 16 *The iterates of Algorithm 2 and 3 run with $\mathcal{C}_i^k \in \mathbb{B}(\alpha_P)$ satisfy*

$$
\mathbb{E}\left[\left\| X_i^{k+1} - W_i^{k+1} \right\|_{(i)}^2 \right] \leq \left(1 - \frac{\alpha_P}{2}\right) \mathbb{E}\left[\left\| X_i^k - W_i^k \right\|_{(i)}^2 \right] + \frac{2}{\alpha_P} (\gamma_i^k)^2 \mathbb{E}\left[\left\| G_i^k \right\|_{(i)\star}^2 \right].
$$

Proof Let $\mathbb{E}_{\mathcal{C}}[\cdot]$ denote the expectation over the randomness introduced by the compressors. Then

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{C}}\left[\left\| X_i^{k+1} - W_i^{k+1} \right\|_{(i)}^2 \right] \\
&= \quad \mathbb{E}_{\mathcal{C}}\left[\left\| W_i^k + \mathcal{C}_i^k (X_i^{k+1} - W_i^k) - X_i^{k+1} \right\|_{(i)}^2 \right]
\end{aligned}
$$

$$\begin{aligned}
&\overset{(1)}{\leq} \quad (1-\alpha_P)\left\|X_i^{k+1}-W_i^k\right\|_{(i)}^2 \\
&\overset{(27)}{\leq} \quad (1-\alpha_P)\left(1+\frac{\alpha_P}{2}\right)\left\|X_i^k-W_i^k\right\|_{(i)}^2+(1-\alpha_P)\left(1+\frac{2}{\alpha_P}\right)\left\|X_i^{k+1}-X_i^k\right\|_{(i)}^2 \\
&\overset{(29),(30)}{\leq} \quad \left(1-\frac{\alpha_P}{2}\right)\left\|X_i^k-W_i^k\right\|_2^2+\frac{2}{\alpha_P}\left\|X_i^{k+1}-X_i^k\right\|_{(i)}^2 .
\end{aligned}$$

It remains to take full expectation and use the fact that

$$\left\|X_i^{k+1}-X_i^k\right\|_{(i)}=\gamma_i^k\left\|\left(G_i^k\right)^\sharp\right\|_{(i)}\overset{(33)}{=}\gamma_i^k\left\|G_i^k\right\|_{(i)\star}.$$

■

F.2.1. SMOOTH CASE

Lemma 17 *Let Assumptions 7 and 10 hold. Then, the iterates of Algorithm 3 run with $\mathcal{C}_{i,j}^k \in \mathbb{B}_2(\alpha_P)$ satisfy*

$$\begin{aligned}
\mathbb{E}\left[\left\|M_{i,j}^{k+1}-G_{i,j}^{k+1}\right\|_2^2\right] \quad \leq \quad & \left(1-\frac{\alpha_D}{2}\right)\mathbb{E}\left[\left\|M_{i,j}^k-G_{i,j}^k\right\|_2^2\right]+\frac{6\beta_i^2}{\alpha_D}\mathbb{E}\left[\left\|M_{i,j}^k-\nabla_i f_j(X^k)\right\|_2^2\right] \\
&+\frac{6\beta_i^2}{\alpha_D\underline{\rho}_i^2}(L_{i,j}^0)^2(\gamma_i^k)^2\mathbb{E}\left[\left\|G_i^k\right\|_\star^2\right] \\
&+\frac{6\beta_i^2}{\alpha_D\underline{\rho}_i^2}(L_{i,j}^0)^2\mathbb{E}\left[\left\|X_i^{k+1}-W_i^{k+1}\right\|_{(i)}^2\right]+(1-\alpha_D)\beta_i^2\sigma_i^2.
\end{aligned}$$

Proof Using the definition of contractive compressors and the algorithm's momentum update rule, we get

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\left[\left\|M_{i,j}^{k+1}-G_{i,j}^{k+1}\right\|_2^2\right] \quad &= \quad \mathbb{E}_{\mathcal{C}}\left[\left\|M_{i,j}^{k+1}-G_{i,j}^k-\mathcal{C}_{i,j}^k(M_{i,j}^{k+1}-G_{i,j}^k)\right\|_2^2\right] \\
&\overset{(1)}{\leq} \quad (1-\alpha_D)\left\|M_{i,j}^{k+1}-G_{i,j}^k\right\|_2^2,
\end{aligned}$$

where $\mathbb{E}_{\mathcal{C}}[\cdot]$ denotes the expectation over the randomness introduced by the compressors. Then, letting $\mathbb{E}_\xi[\cdot]$ be the expectation over the stochasticity of the gradients, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|M_{i,j}^{k+1}-G_{i,j}^{k+1}\right\|_2^2\right] \\
&\leq \quad \mathbb{E}\left[\mathbb{E}_{\mathcal{C}}\left[\left\|M_{i,j}^{k+1}-G_{i,j}^{k+1}\right\|_2^2\right]\right] \\
&\leq \quad (1-\alpha_D)\mathbb{E}\left[\left\|M_{i,j}^{k+1}-G_{i,j}^k\right\|_2^2\right] \\
&= \quad (1-\alpha_D)\mathbb{E}\left[\mathbb{E}_\xi\left[\left\|(1-\beta_i)M_{i,j}^k+\beta_i\nabla_i f_j(W^{k+1},\xi_j^{k+1})-G_{i,j}^k\right\|_2^2\right]\right]
\end{aligned}$$

$$\begin{aligned}
&\overset{(40)}{=} && (1-\alpha_D)\mathbb{E}\left[\left\|(1-\beta_i)M_{i,j}^k + \beta_i\nabla_i f_j(W^{k+1}) - G_{i,j}^k\right\|_2^2\right] \\
& && +(1-\alpha_D)\beta_i^2\mathbb{E}\left[\left\|\nabla_i f_j(W^{k+1},\xi_j^{k+1}) - \nabla_i f_j(W^{k+1})\right\|_2^2\right] \\
&\overset{(27)}{\leq} && (1-\alpha_D)\left(1+\frac{\alpha_D}{2}\right)\mathbb{E}\left[\left\|M_{i,j}^k - G_{i,j}^k\right\|_2^2\right] \\
& && +(1-\alpha_D)\left(1+\frac{2}{\alpha_D}\right)\beta_i^2\mathbb{E}\left[\left\|M_{i,j}^k - \nabla_i f_j(W^{k+1})\right\|_2^2\right] + (1-\alpha_D)\beta_i^2\sigma_i^2,
\end{aligned}$$

where in the last line we used Assumption 10. Then, Assumption 7 gives

$$\begin{aligned}
&\mathbb{E}\left[\left\|M_{i,j}^{k+1} - G_{i,j}^{k+1}\right\|_2^2\right] \\
&\overset{(29),(30)}{\leq} && \left(1-\frac{\alpha_D}{2}\right)\mathbb{E}\left[\left\|M_{i,j}^k - G_{i,j}^k\right\|_2^2\right] + \frac{2}{\alpha_D}\beta_i^2\mathbb{E}\left[\left\|M_{i,j}^k - \nabla_i f_j(W^{k+1})\right\|_2^2\right] + (1-\alpha_D)\beta_i^2\sigma_i^2 \\
&\overset{(27)}{\leq} && \left(1-\frac{\alpha_D}{2}\right)\mathbb{E}\left[\left\|M_{i,j}^k - G_{i,j}^k\right\|_2^2\right] + \frac{6\beta_i^2}{\alpha_D}\mathbb{E}\left[\left\|M_{i,j}^k - \nabla_i f_j(X^k)\right\|_2^2\right] \\
& && +\frac{6\beta_i^2}{\alpha_D}\mathbb{E}\left[\left\|\nabla_i f_j(X^k) - \nabla_i f_j(X^{k+1})\right\|_2^2\right] \\
& && +\frac{6\beta_i^2}{\alpha_D}\mathbb{E}\left[\left\|\nabla_i f_j(X^{k+1}) - \nabla_i f_j(W^{k+1})\right\|_2^2\right] + (1-\alpha_D)\beta_i^2\sigma_i^2 \\
&\leq && \left(1-\frac{\alpha_D}{2}\right)\mathbb{E}\left[\left\|M_{i,j}^k - G_{i,j}^k\right\|_2^2\right] + \frac{6\beta_i^2}{\alpha_D}\mathbb{E}\left[\left\|M_{i,j}^k - \nabla_i f_j(X^k)\right\|_2^2\right] \\
& && +\frac{6\beta_i^2}{\alpha_D\underline{\rho}_i^2}(L_{i,j}^0)^2\mathbb{E}\left[\left\|X_i^k - X_i^{k+1}\right\|_{(i)}^2\right] \\
& && +\frac{6\beta_i^2}{\alpha_D\underline{\rho}_i^2}(L_{i,j}^0)^2\mathbb{E}\left[\left\|X_i^{k+1} - W_i^{k+1}\right\|_{(i)}^2\right] + (1-\alpha_D)\beta_i^2\sigma_i^2.
\end{aligned}$$

Noting that $\left\|X_i^{k+1} - X_i^k\right\|_{(i)} = \gamma_i^k\left\|(G_i^k)^\sharp\right\|_{(i)} \overset{(33)}{=} \gamma_i^k\left\|G_i^k\right\|_{(i)\star}$ finishes the proof. ■

Lemma 18 *Let Assumptions 6, 7 and 10 hold. Then, the iterates of Algorithm 3 satisfy*

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla_i f_j(X^{k+1}) - M_{i,j}^{k+1}\right\|_2^2\right] \\
&\quad\leq \left(1-\frac{\beta_i}{2}\right)\mathbb{E}\left[\left\|\nabla_i f_j(X^k) - M_{i,j}^k\right\|_2^2\right] + \frac{2}{\beta_i\underline{\rho}_i^2}(L_{i,j}^0)^2(\gamma_i^k)^2\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right] \\
&\quad\quad +\frac{\beta_i^2}{\underline{\rho}_i^2}\left(1+\frac{2}{\beta_i}\right)(L_{i,j}^0)^2\mathbb{E}\left[\left\|X_i^{k+1} - W_i^{k+1}\right\|_{(i)}^2\right] + \beta_i^2\sigma_i^2
\end{aligned}$$

and

$$\mathbb{E}\left[\left\|\nabla_i f(X^{k+1}) - M_i^{k+1}\right\|_2^2\right]$$

$$\begin{aligned}
&\leq \left(1-\frac{\beta_i}{2}\right)\left\|\nabla_i f(X^k)-M_i^k\right\|_2^2+\frac{2}{\beta_i\underline{\rho}_i^2}(L_i^0)^2(\gamma_i^k)^2\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right]\\
&\quad+\frac{\beta_i^2}{\underline{\rho}_i^2}\left(1+\frac{2}{\beta_i}\right)(L_i^0)^2\mathbb{E}\left[\left\|X_i^{k+1}-W_i^{k+1}\right\|_{(i)}^2\right]+\frac{\beta_i^2\sigma_i^2}{n},
\end{aligned}$$

where $M_i^k := \frac{1}{n}\sum_{j=1}^n M_{i,j}^k$.

Proof Using the momentum update rule and letting $\mathbb{E}_\xi\left[\cdot\right]$ be the expectation over the stochasticity of the gradients, we get

$$\begin{aligned}
&\mathbb{E}_\xi\left[\left\|\nabla_i f_j(X^{k+1})-M_{i,j}^{k+1}\right\|_2^2\right]\\
&= \mathbb{E}_\xi\left[\left\|\nabla_i f_j(X^{k+1})-(1-\beta_i)M_{i,j}^k-\beta_i\nabla_i f_j(W^{k+1},\xi_j^{k+1})\right\|_2^2\right]\\
&\overset{(40)}{=} \left\|\nabla_i f_j(X^{k+1})-(1-\beta_i)M_{i,j}^k-\beta_i\nabla_i f_j(W^{k+1})\right\|_2^2\\
&\quad+\beta_i^2\mathbb{E}_\xi\left[\left\|\nabla_i f_j(W^{k+1},\xi_j^{k+1})-\nabla_i f_j(W^{k+1})\right\|_2^2\right]\\
&\overset{(27)}{\leq} (1-\beta_i)^2\left(1+\frac{\beta_i}{2}\right)\mathbb{E}_\xi\left[\left\|\nabla_i f_j(X^{k+1})-M_{i,j}^k\right\|_2^2\right]\\
&\quad+\beta_i^2\left(1+\frac{2}{\beta_i}\right)\mathbb{E}_\xi\left[\left\|\nabla_i f_j(X^{k+1})-\nabla_i f_j(W^{k+1})\right\|_2^2\right]\\
&\quad+\beta_i^2\mathbb{E}_\xi\left[\left\|\nabla_i f_j(W^{k+1},\xi_j^{k+1})-\nabla_i f_j(W^{k+1})\right\|_2^2\right]\\
&\overset{(29)}{\leq} (1-\beta_i)\left\|\nabla_i f_j(X^{k+1})-M_{i,j}^k\right\|_2^2\\
&\quad+\beta_i^2\left(1+\frac{2}{\beta_i}\right)\left\|\nabla_i f_j(X^{k+1})-\nabla_i f_j(W^{k+1})\right\|_2^2+\beta_i^2\sigma_i^2,
\end{aligned}$$

where in the last line we used Assumption 5. Then, Assumption 7 gives

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla_i f_j(X^{k+1})-M_{i,j}^{k+1}\right\|_2^2\right]\\
&= \mathbb{E}\left[\mathbb{E}_\xi\left[\left\|\nabla_i f_j(X^{k+1})-M_{i,j}^{k+1}\right\|_2^2\right]\right]\\
&\overset{(27)}{\leq} (1-\beta_i)\left(1+\frac{\beta_i}{2}\right)\mathbb{E}\left[\left\|\nabla_i f_j(X^k)-M_{i,j}^k\right\|_2^2\right]\\
&\quad+(1-\beta_i)\left(1+\frac{2}{\beta_i}\right)\mathbb{E}\left[\left\|\nabla_i f_j(X^{k+1})-\nabla_i f_j(X^k)\right\|_2^2\right]\\
&\quad+\beta_i^2\left(1+\frac{2}{\beta_i}\right)\mathbb{E}\left[\left\|\nabla_i f_j(X^{k+1})-\nabla_i f_j(W^{k+1})\right\|_2^2\right]+\beta_i^2\sigma_i^2\\
&\overset{(29),(30)}{\leq} \left(1-\frac{\beta_i}{2}\right)\mathbb{E}\left[\left\|\nabla_i f_j(X^k)-M_{i,j}^k\right\|_2^2\right]+\frac{2}{\beta_i\underline{\rho}_i^2}(L_{i,j}^0)^2\mathbb{E}\left[\left\|X_i^{k+1}-X_i^k\right\|_{(i)}^2\right]
\end{aligned}$$

$$+\frac{\beta_i^2}{\underline{\rho}_i^2}\left(1+\frac{2}{\beta_i}\right)(L^0_{i,j})^2\mathbb{E}\left[\left\|X_i^{k+1}-W_i^{k+1}\right\|^2_{(i)}\right]+\beta_i^2\sigma_i^2.$$

To prove the second part of the statement, define $\nabla_i f(X,\xi^k) := \frac{1}{n}\sum_{i=1}^n \nabla_i f_j(X,\xi_j^k)$. Then $M_i^{k+1} = (1-\beta_i)M_i^k + \beta_i \nabla_i f(W^{k+1},\xi^{k+1})$, so following similar steps as above, we get

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla_i f(X^{k+1}) - M_i^{k+1}\right\|_2^2\right] \\
&= \mathbb{E}\left[\mathbb{E}_\xi\left[\left\|\nabla_i f(X^{k+1}) - (1-\beta_i)M_i^k - \beta_i\nabla_i f(W^{k+1},\xi^{k+1})\right\|_2^2\right]\right] \\
&\overset{(40)}{=} \mathbb{E}\left[\left\|\nabla_i f(X^{k+1}) - (1-\beta_i)M_i^k - \beta_i\nabla_i f(W^{k+1})\right\|_2^2\right] \\
&\quad +\beta_i^2\mathbb{E}\left[\mathbb{E}_\xi\left[\left\|\nabla_i f(W^{k+1},\xi^{k+1}) - \nabla_i f(W^{k+1})\right\|_2^2\right]\right] \\
&\overset{(27)}{\leq} (1-\beta_i)^2\left(1+\frac{\beta_i}{2}\right)\mathbb{E}\left[\left\|\nabla_i f(X^{k+1}) - M_i^k\right\|_2^2\right] \\
&\quad +\beta_i^2\left(1+\frac{2}{\beta_i}\right)\mathbb{E}\left[\left\|\nabla_i f(X^{k+1}) - \nabla_i f(W^{k+1})\right\|_2^2\right] \\
&\quad +\beta_i^2\mathbb{E}\left[\mathbb{E}_\xi\left[\left\|\nabla_i f(W^{k+1},\xi^{k+1}) - \nabla_i f(W^{k+1})\right\|_2^2\right]\right] \\
&\overset{(29)}{\leq} (1-\beta_i)\mathbb{E}\left[\left\|\nabla_i f(X^{k+1}) - M_i^k\right\|_2^2\right] \\
&\quad +\beta_i^2\left(1+\frac{2}{\beta_i}\right)\mathbb{E}\left[\left\|\nabla_i f(X^{k+1}) - \nabla_i f(W^{k+1})\right\|_2^2\right] + \frac{\beta_i^2\sigma_i^2}{n} \\
&\overset{(27)}{\leq} (1-\beta_i)\left(1+\frac{\beta_i}{2}\right)\mathbb{E}\left[\left\|\nabla_i f(X^k) - M_i^k\right\|_2^2\right] \\
&\quad +(1-\beta_i)\left(1+\frac{2}{\beta_i}\right)\mathbb{E}\left[\left\|\nabla_i f(X^{k+1}) - \nabla_i f(X^k)\right\|_2^2\right] \\
&\quad +\beta_i^2\left(1+\frac{2}{\beta_i}\right)\mathbb{E}\left[\left\|\nabla_i f(X^{k+1}) - \nabla_i f(W^{k+1})\right\|_2^2\right] + \frac{\beta_i^2\sigma_i^2}{n} \\
&\overset{(29),(30)}{\leq} \left(1-\frac{\beta_i}{2}\right)\mathbb{E}\left[\left\|\nabla_i f(X^k) - M_i^k\right\|_2^2\right] + \frac{2}{\beta_i\underline{\rho}_i^2}(L_i^0)^2\mathbb{E}\left[\left\|X_i^{k+1}-X_i^k\right\|^2_{(i)}\right] \\
&\quad +\frac{\beta_i^2}{\underline{\rho}_i^2}\left(1+\frac{2}{\beta_i}\right)(L_i^0)^2\mathbb{E}\left[\left\|X_i^{k+1}-W_i^{k+1}\right\|^2_{(i)}\right] + \frac{\beta_i^2\sigma_i^2}{n}.
\end{aligned}$$

It remains to use the fact that $\left\|X_i^{k+1}-X_i^k\right\|_{(i)} = \gamma_i^k\left\|(G_i^k)^\sharp\right\|_{(i)} \overset{(33)}{=} \gamma_i^k\left\|G_i^k\right\|_{(i)\star}$. ■

F.2.2. GENERALIZED SMOOTH CASE

Lemma 19 *Let Assumption 9 hold. Then, the iterates of Algorithm 2 run with $\mathcal{C}_i^k \equiv \mathcal{I}$ (the identity compressor) and $\mathcal{C}_{i,j}^k \in \mathbb{B}_\star(\alpha_D)$ satisfy*

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla_i f_j(X^{k+1}) - G_{i,j}^{k+1}\right\|_{(i)\star}\middle| X^{k+1}, G^k\right] \\
&\quad \leq \sqrt{1-\alpha_D}\left\|\nabla_i f_j(X^k) - G_{i,j}^k\right\|_{(i)\star} + \sqrt{1-\alpha_D}\left(L_{i,j}^0 + L_{i,j}^1\left\|\nabla_i f_j(X^k)\right\|_{(i)\star}\right)t_i^k.
\end{aligned}$$

Proof The algorithm's update rule and Jensen's inequality give

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla_i f_j(X^{k+1}) - G_{i,j}^{k+1}\right\|_{(i)\star}\middle| X^{k+1}, G^k\right] \\
&\quad = \mathbb{E}\left[\sqrt{\left\|\nabla_i f_j(X^{k+1}) - G_{i,j}^k - \mathcal{C}_{i,j}^k(\nabla_i f_j(X^{k+1}) - G_{i,j}^k)\right\|_{(i)\star}^2}\middle| X^{k+1}, G^k\right] \\
&\quad \leq \sqrt{\mathbb{E}\left[\left\|\nabla_i f_j(X^{k+1}) - G_{i,j}^k - \mathcal{C}_{i,j}^k(\nabla_i f_j(X^{k+1}) - G_{i,j}^k)\right\|_{(i)\star}^2\middle| X^{k+1}, G^k\right]} \\
&\quad \leq \sqrt{1-\alpha_D}\left\|\nabla_i f_j(X^{k+1}) - G_{i,j}^k\right\|_{(i)\star} \\
&\quad \leq \sqrt{1-\alpha_D}\left\|\nabla_i f_j(X^k) - G_{i,j}^k\right\|_{(i)\star} + \sqrt{1-\alpha_D}\left\|\nabla_i f_j(X^{k+1}) - \nabla_i f_j(X^k)\right\|_{(i)\star} \\
&\quad \leq \sqrt{1-\alpha_D}\left\|\nabla_i f_j(X^k) - G_{i,j}^k\right\|_{(i)\star} \\
&\qquad + \sqrt{1-\alpha_D}\left(L_{i,j}^0 + L_{i,j}^1\left\|\nabla_i f_j(X^k)\right\|_{(i)\star}\right)\left\|X_i^{k+1} - X_i^k\right\|_{(i)}.
\end{aligned}$$

where $\left\|X_i^{k+1} - X_i^k\right\|_{(i)} = t_i^k$. ■

Lemma 20 *Let Assumptions 9 and 10 hold. Then, the iterates of Algorithm 3 run with $\mathcal{C}_i^k \equiv \mathcal{I}$ (the identity compressor) and $\mathcal{C}_{i,j}^k \in \mathbb{B}_2(\alpha_D)$ satisfy*

$$\begin{aligned}
&\mathbb{E}\left[\left\|M_{i,j}^{k+1} - G_{i,j}^{k+1}\right\|_2\middle| X^{k+1}, M_{i,j}^k, G_{i,j}^k\right] \\
&\quad \leq \sqrt{1-\alpha_D}\left\|M_{i,j}^k - G_{i,j}^k\right\|_2 + \sqrt{1-\alpha_D}\beta_i\left\|M_{i,j}^k - \nabla_i f_j(X^k)\right\|_2 \\
&\qquad + \frac{\sqrt{1-\alpha_D}\beta_i}{\underline{\rho}_i}\left(L_{i,j}^0 + L_{i,j}^1\left\|\nabla_i f_j(X^k)\right\|_{(i)\star}\right)t_i^k + \sqrt{1-\alpha_D}\beta_i\sigma_i.
\end{aligned}$$

Proof Using the definition of contractive compressors and triangle inequality, we get

$$\begin{aligned}
&\mathbb{E}\left[\left\|M_{i,j}^{k+1} - G_{i,j}^{k+1}\right\|_2\middle| M_{i,j}^{k+1}, G_{i,j}^k\right] \\
&\quad = \mathbb{E}\left[\sqrt{\left\|M_{i,j}^{k+1} - G_{i,j}^k - \mathcal{C}_{i,j}^k(M_{i,j}^{k+1} - G_{i,j}^k)\right\|_2^2}\middle| M_{i,j}^{k+1}, G_{i,j}^k\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \sqrt{\mathbb{E}\left[\left\|M_{i,j}^{k+1} - G_{i,j}^{k} - \mathcal{C}_{i,j}^{k}(M_{i,j}^{k+1} - G_{i,j}^{k})\right\|_2^2 \,\middle|\, M_{i,j}^{k+1}, G_{i,j}^{k}\right]} \\
&\overset{(1)}{\leq} \sqrt{1-\alpha_D}\left\|M_{i,j}^{k+1} - G_{i,j}^{k}\right\|_2 \\
&= \sqrt{1-\alpha_D}\left\|(1-\beta_i)M_{i,j}^{k} + \beta_i \nabla_i f_j(X^{k+1}, \xi_j^{k+1}) - G_{i,j}^{k}\right\|_2 .
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|M_{i,j}^{k+1} - G_{i,j}^{k+1}\right\|_2 \,\middle|\, X^{k+1}, M_{i,j}^{k}, G_{i,j}^{k}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\|M_{i,j}^{k+1} - G_{i,j}^{k+1}\right\|_2 \,\middle|\, M_{i,j}^{k+1}, G_{i,j}^{k}\right] \,\middle|\, X^{k+1}, M_{i,j}^{k}, G_{i,j}^{k}\right] \\
&\leq \sqrt{1-\alpha_D}\,\mathbb{E}\left[\left\|(1-\beta_i)M_{i,j}^{k} + \beta_i \nabla_i f_j(X^{k+1}, \xi_j^{k+1}) - G_{i,j}^{k}\right\|_2 \,\middle|\, X^{k+1}, M_{i,j}^{k}, G_{i,j}^{k}\right] \\
&\leq \sqrt{1-\alpha_D}\,\mathbb{E}\left[\left\|(1-\beta_i)M_{i,j}^{k} + \beta_i \nabla_i f_j(X^{k+1}) - G_{i,j}^{k}\right\|_2 \,\middle|\, X^{k+1}, M_{i,j}^{k}, G_{i,j}^{k}\right] \\
&\quad + \sqrt{1-\alpha_D}\,\beta_i\,\mathbb{E}\left[\left\|\nabla_i f_j(X^{k+1}, \xi_j^{k+1}) - \nabla_i f_j(X^{k+1})\right\|_2 \,\middle|\, X^{k+1}, M_{i,j}^{k}, G_{i,j}^{k}\right] \\
&\overset{(10)}{\leq} \sqrt{1-\alpha_D}\left\|M_{i,j}^{k} - G_{i,j}^{k}\right\|_2 + \sqrt{1-\alpha_D}\,\beta_i\left\|M_{i,j}^{k} - \nabla_i f_j(X^{k+1})\right\|_2 + \sqrt{1-\alpha_D}\,\beta_i\sigma_i \\
&\leq \sqrt{1-\alpha_D}\left\|M_{i,j}^{k} - G_{i,j}^{k}\right\|_2 + \sqrt{1-\alpha_D}\,\beta_i\left\|M_{i,j}^{k} - \nabla_i f_j(X^{k})\right\|_2 \\
&\quad + \sqrt{1-\alpha_D}\,\beta_i\left\|\nabla_i f_j(X^{k}) - \nabla_i f_j(X^{k+1})\right\|_2 + \sqrt{1-\alpha_D}\,\beta_i\sigma_i \\
&\overset{(9)}{\leq} \sqrt{1-\alpha_D}\left\|M_{i,j}^{k} - G_{i,j}^{k}\right\|_2 + \sqrt{1-\alpha_D}\,\beta_i\left\|M_{i,j}^{k} - \nabla_i f_j(X^{k})\right\|_2 \\
&\quad + \frac{\sqrt{1-\alpha_D}\,\beta_i}{\underline{\rho}_i}\left(L_{i,j}^{0} + L_{i,j}^{1}\left\|\nabla_i f_j(X^{k})\right\|_{(i)\star}\right)\left\|X_i^{k} - X_i^{k+1}\right\|_{(i)} + \sqrt{1-\alpha_D}\,\beta_i\sigma_i .
\end{aligned}
$$

Using the fact that $\left\|X_i^{k} - X_i^{k+1}\right\| = t_i^{k}$ finishes the proof. ■

Lemma 21 *Let Assumptions 8, 9 and 10 hold. Then, the iterates of Algorithm 3 run with $\mathcal{C}_i^k \equiv \mathcal{I}$ (the identity compressor) satisfy*

$$
\begin{aligned}
\mathbb{E}\left[\left\|M_i^{k+1} - \nabla_i f(X^{k+1})\right\|_2\right] &\leq (1-\beta_i)^{k+1}\,\mathbb{E}\left[\left\|M_i^{0} - \nabla_i f(X^{0})\right\|_2\right] + \frac{t_i^k \bar{L}_i^0}{\beta_i \underline{\rho}_i} \\
&\quad + \frac{t_i^k}{\underline{\rho}_i}\frac{1}{n}\sum_{j=1}^{n} L_{i,j}^{1}\sum_{l=0}^{k}(1-\beta_i)^{k+1-l}\,\mathbb{E}\left[\left\|\nabla_i f_j(X^{l})\right\|_{(i)\star}\right] + \sigma_i\sqrt{\frac{\beta_i}{n}}
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|M_{i,j}^{k+1} - \nabla_i f_j(X^{k+1})\right\|_2\right] &\leq (1-\beta_i)\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|M_{i,j}^{k} - \nabla_i f_j(X^{k})\right\|_2\right] + t_i^k\frac{(1-\beta_i)\bar{L}_i^0}{\underline{\rho}_i} \\
&\quad + t_i^k\frac{1-\beta_i}{\underline{\rho}_i}\frac{1}{n}\sum_{j=1}^{n} L_{i,j}^{1}\,\mathbb{E}\left[\left\|\nabla_i f_j(X^{k})\right\|_{(i)\star}\right] + \beta_i\sigma_i,
\end{aligned}
$$

where $M_i^k := \frac{1}{n}\sum_{j=1}^n M_{i,j}^k$.

Proof The proof uses techniques similar to those in Cutkosky and Mehta [10, Theorem 1]. First, using the momentum update rule, we can write

$$\begin{aligned} M_{i,j}^{k+1} &= (1-\beta_i)M_{i,j}^k + \beta_i \nabla_i f_j(X^{k+1}, \xi_j^{k+1}) \\ &= (1-\beta_i)\left(M_{i,j}^k - \nabla_i f_j(X^k)\right) + (1-\beta_i)\left(\nabla_i f_j(X^k) - \nabla_i f_j(X^{k+1})\right) \\ &\quad +\beta_i\left(\nabla_i f_j(X^{k+1}, \xi_j^{k+1}) - \nabla_i f_j(X^{k+1})\right) + \nabla_i f_j(X^{k+1}), \end{aligned}$$

and hence

$$U_{1,i,j}^{k+1} = (1-\beta_i)U_{1,i,j}^k + (1-\beta_i)U_{2,i,j}^k + \beta_i U_{3,i,j}^{k+1},$$

where we define $U_{1,i,j}^k := M_{i,j}^k - \nabla_i f_j(X^k)$, $U_{2,i,j}^k := \nabla_i f_j(X^k) - \nabla_i f_j(X^{k+1})$ and $U_{3,i,j}^k := \nabla_i f_j(X^k, \xi_j^k) - \nabla_i f_j(X^k)$. Unrolling the recursion gives

$$U_{1,i,j}^{k+1} = (1-\beta_i)^{k+1}U_{1,i,j}^0 + \sum_{l=0}^k (1-\beta_i)^{k+1-l}U_{2,i,j}^l + \beta_i \sum_{l=0}^k (1-\beta_i)^{k-l}U_{3,i,j}^{l+1}.$$

Hence, using the triangle inequality,

$$\begin{aligned} &\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^n U_{1,i,j}^{k+1}\right\|_2\right] \\ &\leq (1-\beta_i)^{k+1}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^n U_{1,i,j}^0\right\|_2\right] + \mathbb{E}\left[\left\|\sum_{l=0}^k (1-\beta_i)^{k+1-l}\frac{1}{n}\sum_{j=1}^n U_{2,i,j}^l\right\|_2\right] \\ &\quad +\beta_i \mathbb{E}\left[\left\|\sum_{l=0}^k (1-\beta_i)^{k-l}\frac{1}{n}\sum_{j=1}^n U_{3,i,j}^{l+1}\right\|_2\right]. \end{aligned} \tag{10}$$

Let us now bound the last two terms of the inequality above. First, triangle inequality and Assumption 9 give

$$\begin{aligned} &\mathbb{E}\left[\left\|\sum_{l=0}^k (1-\beta_i)^{k+1-l}\frac{1}{n}\sum_{j=1}^n U_{2,i,j}^l\right\|_2\right] \\ &\leq \frac{1}{n}\sum_{j=1}^n\sum_{l=0}^k (1-\beta_i)^{k+1-l}\mathbb{E}\left[\left\|U_{2,i,j}^l\right\|_2\right] \\ &= \frac{1}{n}\sum_{j=1}^n\sum_{l=0}^k (1-\beta_i)^{k+1-l}\mathbb{E}\left[\left\|\nabla_i f_j(X^l) - \nabla_i f_j(X^{l+1})\right\|_2\right] \\ &\overset{(9)}{\leq} \frac{1}{\underline{\rho}_i}\frac{1}{n}\sum_{j=1}^n\sum_{l=0}^k (1-\beta_i)^{k+1-l}\mathbb{E}\left[\left(L_{i,j}^0 + L_{i,j}^1\left\|\nabla_i f_j(X^l)\right\|_{(i)\star}\right)\left\|X_i^l - X_i^{l+1}\right\|_{(i)}\right] \end{aligned}$$

$$
\begin{aligned}
&= \frac{t_i^k}{\underline{\rho}_i}\frac{1}{n}\sum_{j=1}^n\sum_{l=0}^k(1-\beta_i)^{k+1-l}L_{i,j}^0 + \frac{t_i^k}{\underline{\rho}_i}\frac{1}{n}\sum_{j=1}^n L_{i,j}^1\sum_{l=0}^k(1-\beta_i)^{k+1-l}\mathbb{E}\left[\left\|\nabla_i f_j(X^l)\right\|_{(i)\star}\right] \\
&\le \frac{t_i^k\bar{L}_i^0}{\beta_i\underline{\rho}_i} + \frac{t_i^k}{\underline{\rho}_i}\frac{1}{n}\sum_{j=1}^n L_{i,j}^1\sum_{l=0}^k(1-\beta_i)^{k+1-l}\mathbb{E}\left[\left\|\nabla_i f_j(X^l)\right\|_{(i)\star}\right],
\end{aligned}
$$

and using Jensen's inequality, the last term can be bounded as

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{l=0}^k(1-\beta_i)^{k-l}\frac{1}{n}\sum_{j=1}^n U_{3,i,j}^{l+1}\right\|_2\right] \\
&\le \sqrt{\mathbb{E}\left[\left\|\sum_{l=0}^k(1-\beta_i)^{k-l}\frac{1}{n}\sum_{j=1}^n U_{3,i,j}^{l+1}\right\|_2^2\right]} \overset{(10)}{=} \sqrt{\sum_{l=0}^k(1-\beta_i)^{2(k-l)}\frac{1}{n^2}\sum_{j=1}^n\mathbb{E}\left[\left\|U_{3,i,j}^{l+1}\right\|_2^2\right]} \\
&\overset{(10)}{\le} \sqrt{\sum_{l=0}^k(1-\beta_i)^{2(k-l)}\frac{1}{n^2}\sum_{j=1}^n\sigma_i^2} = \frac{\sigma_i}{\sqrt{n}}\sqrt{\sum_{l=0}^k(1-\beta_i)^{2l}} \le \frac{\sigma_i}{\sqrt{n\beta_i(2-\beta_i)}} \le \frac{\sigma_i}{\sqrt{n\beta_i}}.
\end{aligned}
$$

Substituting this in (10) yields

$$
\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^n U_{1,i,j}^{k+1}\right\|_2\right] &\le (1-\beta_i)^{k+1}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^n U_{1,i,j}^0\right\|_2\right] + \frac{t_i\bar{L}_i^0}{\beta_i\underline{\rho}_i} \\
&\quad + \frac{t_i^k}{\underline{\rho}_i}\frac{1}{n}\sum_{j=1}^n L_{i,j}^1\sum_{l=0}^k(1-\beta_i)^{k+1-l}\mathbb{E}\left[\left\|\nabla_i f_j(X^l)\right\|_{(i)\star}\right] + \beta_i\frac{\sigma_i}{\sqrt{n\beta_i}}.
\end{aligned}
$$

To prove the second inequality, recall that $U_{1,i,j}^{k+1} = (1-\beta_i)U_{1,i,j}^k + (1-\beta_i)U_{2,i,j}^k + \beta_i U_{3,i,j}^{k+1}$. Hence, taking norms, averaging, and using the triangle inequality,

$$
\begin{aligned}
\frac{1}{n}\sum_{j=1}^n\mathbb{E}\left[\left\|U_{1,i,j}^{k+1}\right\|_2\right] &\le (1-\beta_i)\frac{1}{n}\sum_{j=1}^n\mathbb{E}\left[\left\|U_{1,i,j}^k\right\|_2\right] + (1-\beta_i)\frac{1}{n}\sum_{j=1}^n\mathbb{E}\left[\left\|U_{2,i,j}^k\right\|_2\right] \\
&\quad + \beta_i\frac{1}{n}\sum_{j=1}^n\mathbb{E}\left[\left\|U_{3,i,j}^{k+1}\right\|_2\right],
\end{aligned} \tag{11}
$$

where the last two terms can be bounded as

$$
\begin{aligned}
\frac{1}{n}\sum_{j=1}^n\mathbb{E}\left[\left\|U_{2,i,j}^k\right\|_2\right] &= \frac{1}{n}\sum_{j=1}^n\mathbb{E}\left[\left\|\nabla_i f_j(X^k) - \nabla_i f_j(X^{k+1})\right\|_2\right] \\
&\overset{(9)}{\le} \frac{1}{\underline{\rho}_i}\frac{1}{n}\sum_{j=1}^n\mathbb{E}\left[\left(L_{i,j}^0 + L_{i,j}^1\left\|\nabla_i f_j(X^k)\right\|_{(i)\star}\right)\left\|X_i^k - X_i^{k+1}\right\|_{(i)}\right] \\
&= t_i^k\frac{\bar{L}_i^0}{\underline{\rho}_i} + \frac{t_i^k}{\underline{\rho}_i}\frac{1}{n}\sum_{j=1}^n L_{i,j}^1\mathbb{E}\left[\left\|\nabla_i f_j(X^k)\right\|_{(i)\star}\right]
\end{aligned}
$$

and

$$\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|U_{3,i,j}^{k+1}\right\|_2\right] = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|\nabla_i f_j(X^{k+1},\xi_j^k) - \nabla_i f_j(X^{k+1})\right\|_2\right] \overset{(10)}{\leq} \sigma_i.$$

It remains to substitute this in (11) to obtain

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|U_{1,i,j}^{k+1}\right\|_2\right] &\leq (1-\beta_i)\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|U_{1,i,j}^{k}\right\|_2\right] + t_i^k\frac{(1-\beta_i)\bar{L}_i^0}{\underline{\rho}_i} \\ &\quad + t_i^k\frac{1-\beta_i}{\underline{\rho}_i}\frac{1}{n}\sum_{j=1}^{n}L_{i,j}^1\mathbb{E}\left[\left\|\nabla_i f_j(X^k)\right\|_{(i)\star}\right] + \beta_i\sigma_i.\end{aligned}$$

■

Lemma 22 *Let Assumptions 1 and 8 hold. Then*

$$\sum_{i=1}^{p}\frac{\|\nabla_i f(X)\|_{(i)\star}^2}{2\left(L_i^0 + L_i^1\|\nabla_i f(X)\|_{(i)\star}\right)} \leq f(X) - f^\star$$

for any $X = [X_1,\ldots,X_p] \in \mathcal{S}$.

Proof Let $Y = [Y_1,\ldots,Y_p] \in \mathcal{S}$, where $Y_i = X_i - \frac{\|\nabla_i f(X)\|_{(i)\star}}{L_i^0 + L_i^1\|\nabla_i f(X)\|_{(i)\star}}H_i$ for some $H_i \in \partial\|\cdot\|_{(i)\star}(\nabla_i f(X))$. Then, Theorem 41 and the definition of subdifferential give

$$\begin{aligned} f(Y) &\leq f(X) + \langle\nabla f(X), Y-X\rangle + \sum_{i=1}^{p}\frac{L_i^0 + L_i^1\|\nabla_i f(X)\|_{(i)\star}}{2}\|X_i - Y_i\|_{(i)}^2 \\ &= f(X) + \sum_{i=1}^{p}\langle\nabla_i f(X), Y_i - X_i\rangle_{(i)} + \sum_{i=1}^{p}\frac{L_i^0 + L_i^1\|\nabla_i f(X)\|_{(i)\star}}{2}\|X_i - Y_i\|_{(i)}^2 \\ &= f(X) - \sum_{i=1}^{p}\frac{\|\nabla_i f(X)\|_{(i)\star}}{L_i^0 + L_i^1\|\nabla_i f(X)\|_{(i)\star}}\langle\nabla_i f(X), H_i\rangle_{(i)} \\ &\quad + \sum_{i=1}^{p}\left(\frac{L_i^0 + L_i^1\|\nabla_i f(X)\|_{(i)\star}}{2}\frac{\|\nabla_i f(X)\|_{(i)\star}^2}{\left(L_i^0 + L_i^1\|\nabla_i f(X)\|_{(i)\star}\right)^2}\|H_i\|_{(i)}^2\right) \\ &\overset{(7)}{=} f(X) + \sum_{i=1}^{p}\left(-\frac{\|\nabla_i f(X)\|_{(i)\star}^2}{L_i^0 + L_i^1\|\nabla_i f(X)\|_{(i)\star}} + \frac{\|\nabla_i f(X)\|_{(i)\star}^2}{2\left(L_i^0 + L_i^1\|\nabla_i f(X)\|_{(i)\star}\right)}\right) \\ &= f(X) - \sum_{i=1}^{p}\frac{\|\nabla_i f(X)\|_{(i)\star}^2}{2\left(L_i^0 + L_i^1\|\nabla_i f(X)\|_{(i)\star}\right)}, \end{aligned}$$

and hence

$$\sum_{i=1}^{p} \frac{\|\nabla_i f(X)\|_{(i)\star}^2}{2\left(L_i^0 + L_i^1 \|\nabla_i f(X)\|_{(i)\star}\right)} \leq f(X) - f(Y) \leq f(X) - f^\star$$

as needed. ■

Lemma 23 *Let Assumptions 1 and 8 hold. Then, for any $x_i > 0$, $i \in [p]$, we have*

$$\sum_{i=1}^{p} x_i \|\nabla_i f(X)\|_{(i)\star} \leq 4 \max_{i\in[p]}(x_i L_i^1) \left(f(X) - f^\star\right) + \frac{\sum_{i=1}^p x_i^2 L_i^0}{\max_{i\in[p]}(x_i L_i^1)}$$

for all $X \in \mathcal{S}$.

Proof We follow an approach similar to that in Khirirat et al. [29, Lemma 2]. Applying Theorem 22 and Theorem 39 with $y_i = \|\nabla_i f(X)\|_{(i)\star}$, $z_i = L_i^0 + L_i^1 \|\nabla_i f(X)\|_{(i)\star}$ and any positive x_i, we have

$$\begin{aligned}
2\left(f(X) - f^\star\right) &\geq \sum_{i=1}^{p} \frac{\|\nabla_i f(X)\|_{(i)\star}^2}{L_i^0 + L_i^1 \|\nabla_i f(X)\|_{(i)\star}} \\
&\geq \frac{\left(\sum_{i=1}^p x_i \|\nabla_i f(X)\|_{(i)\star}\right)^2}{\sum_{i=1}^p x_i^2 L_i^0 + \sum_{i=1}^p x_i^2 L_i^1 \|\nabla_i f(X)\|_{(i)\star}} \\
&\geq \frac{\left(\sum_{i=1}^p x_i \|\nabla_i f(X)\|_{(i)\star}\right)^2}{\sum_{i=1}^p x_i^2 L_i^0 + \max_{i\in[p]}(x_i L_i^1) \sum_{i=1}^p x_i \|\nabla_i f(X)\|_{(i)\star}} \\
&\geq \begin{cases} \frac{\left(\sum_{i=1}^p x_i \|\nabla_i f(X)\|_{(i)\star}\right)^2}{2\sum_{i=1}^p x_i^2 L_i^0} & \text{if } \frac{\sum_{i=1}^p x_i^2 L_i^0}{\max_{i\in[p]}(x_i L_i^1)} \geq \sum_{i=1}^p x_i \|\nabla_i f(X)\|_{(i)\star}, \\ \frac{\sum_{i=1}^p x_i \|\nabla_i f(X)\|_{(i)\star}}{2\max_{i\in[p]}(x_i L_i^1)} & \text{otherwise.} \end{cases}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sum_{i=1}^{p} x_i \|\nabla_i f(X)\|_{(i)\star} &\leq \max\left\{4\max_{i\in[p]}(x_i L_i^1)\left(f(X) - f^\star\right), \frac{\sum_{i=1}^p x_i^2 L_i^0}{\max_{i\in[p]}(x_i L_i^1)}\right\} \\
&\leq 4\max_{i\in[p]}(x_i L_i^1)\left(f(X) - f^\star\right) + \frac{\sum_{i=1}^p x_i^2 L_i^0}{\max_{i\in[p]}(x_i L_i^1)}.
\end{aligned}$$

■

Lemma 24 *Let Assumptions 1, 2 and 9 hold. Then, for any $x_i > 0$, $i \in [p]$, we have*

$$\begin{aligned}
\sum_{i=1}^{p} x_i \|\nabla_i f_j(X)\|_{(i)\star} &\leq 4\max_{i\in[p]}(x_i L_{i,j}^1)\left(f_j(X) - f^\star\right) + 4\max_{i\in[p]}(x_i L_{i,j}^1)\left(f^\star - f_j^\star\right) \\
&\quad + \frac{\sum_{i=1}^p x_i^2 L_{i,j}^0}{\max_{i\in[p]}(x_i L_{i,j}^1)}
\end{aligned}$$

for all $X \in \mathcal{S}$.

Proof The proof is similar to that of Theorem 23. Applying Theorem 22 and Theorem 39 with $y_i = \left\|\nabla_i f_j(X)\right\|_{(i)\star}$, $z_i = L^0_{i,j} + L^1_{i,j}\left\|\nabla_i f_j(X)\right\|_{(i)\star}$ and any positive x_i, we have

$$\begin{aligned}
2\left(f_j(X) - f_j^\star\right) &\geq \sum_{i=1}^p \frac{\left\|\nabla_i f_j(X)\right\|^2_{(i)\star}}{L^0_{i,j} + L^1_{i,j}\left\|\nabla_i f_j(X)\right\|_{(i)\star}} \\
&\geq \frac{\left(\sum_{i=1}^p x_i\left\|\nabla_i f_j(X)\right\|_{(i)\star}\right)^2}{\sum_{i=1}^p x_i^2 L^0_{i,j} + \sum_{i=1}^p x_i^2 L^1_{i,j}\left\|\nabla_i f_j(X)\right\|_{(i)\star}} \\
&\geq \frac{\left(\sum_{i=1}^p x_i\left\|\nabla_i f_j(X)\right\|_{(i)\star}\right)^2}{\sum_{i=1}^p x_i^2 L^0_{i,j} + \max_{i\in[p]}(x_i L^1_{i,j})\sum_{i=1}^p x_i\left\|\nabla_i f_j(X)\right\|_{(i)\star}} \\
&\geq \begin{cases} \frac{\left(\sum_{i=1}^p x_i\left\|\nabla_i f_j(X)\right\|_{(i)\star}\right)^2}{2\sum_{i=1}^p x_i^2 L^0_{i,j}} & \text{if } \frac{\sum_{i=1}^p x_i^2 L^0_{i,j}}{\max_{i\in[p]}(x_i L^1_{i,j})} \geq \sum_{i=1}^p x_i\left\|\nabla_i f_j(X)\right\|_{(i)\star}, \\ \frac{\sum_{i=1}^p x_i\left\|\nabla_i f_j(X)\right\|_{(i)\star}}{2\max_{i\in[p]}(x_i L^1_{i,j})} & \text{otherwise.} \end{cases}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sum_{i=1}^p x_i\left\|\nabla_i f_j(X)\right\|_{(i)\star} &\leq \max\left\{4\max_{i\in[p]}(x_i L^1_{i,j})\left(f_j(X) - f_j^\star\right), \frac{\sum_{i=1}^p x_i^2 L^0_{i,j}}{\max_{i\in[p]}(x_i L^1_{i,j})}\right\} \\
&\leq 4\max_{i\in[p]}(x_i L^1_{i,j})\left(f_j(X) - f_j^\star\right) + \frac{\sum_{i=1}^p x_i^2 L^0_{i,j}}{\max_{i\in[p]}(x_i L^1_{i,j})} \\
&= 4\max_{i\in[p]}(x_i L^1_{i,j})\left(f_j(X) - f^\star\right) + 4\max_{i\in[p]}(x_i L^1_{i,j})\left(f^\star - f_j^\star\right) \\
&\quad + \frac{\sum_{i=1}^p x_i^2 L^0_{i,j}}{\max_{i\in[p]}(x_i L^1_{i,j})}.
\end{aligned}$$

■

F.3. Deterministic Setting

F.3.1. LAYER-WISE SMOOTH REGIME

Theorem 25 *Let Assumptions 1, 6 and 7 hold. Let $\{X^k\}_{k=0}^{K-1}$, $K \geq 1$, be the iterates of Algorithm 2 run with $\mathcal{C}_i^k \in \mathbb{B}(\alpha_P)$, $\mathcal{C}_{i,j}^k \in \mathbb{B}_\star(\alpha_D)$, and*

$$0 \leq \gamma_i^k \equiv \gamma_i \leq \frac{1}{2L_i^0 + \frac{4}{\alpha_D}\sqrt{12 + \frac{66}{\alpha_P^2}}\tilde{L}_i^0}, \qquad i = 1, \ldots, p.$$

Then

$$\frac{1}{K}\sum_{k=0}^{K-1}\sum_{i=1}^p \frac{\gamma_i}{\frac{1}{p}\sum_{l=1}^p \gamma_l}\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|^2_{(i)\star}\right] \leq \frac{1}{K}\frac{4\Psi^0}{\frac{1}{p}\sum_{l=1}^p \gamma_l},$$

where

$$\begin{aligned}\Psi^k \quad &:= \quad f(X^k) - f^\star + \sum_{i=1}^p \frac{6\gamma_i}{\alpha_D}\frac{1}{n}\sum_{j=1}^n \left\|\nabla_i f_j(X^k) - G_{i,j}^k\right\|_{(i)\star}^2 \\ &\quad + \sum_{i=1}^p \frac{66\gamma_i}{\alpha_D^2}\left(\frac{2}{\alpha_P} - 1\right)(\tilde{L}_i^0)^2 \left\|X_i^k - W_i^k\right\|_{(i)}^2 .\end{aligned}$$

Remark 26 *Theorem 3 follows as a corollary of the more general result above by setting $p = 1$ and initializing with $X^0 = W^0$ and $G_j^0 = \nabla f_j(X^0)$ for all $j \in [n]$.*

Remark 27 *In the Euclidean case and when $p = 1$, our convergence guarantees recover several existing results. When primal compression is disabled (i.e., $\alpha_P = 1$), they match the rate of Richtárik et al. [63, Theorem 1], up to constant factors. With primal compression, the rate coincides with that of* EF21-BC *in Fatkhullin et al. [12, Theorem 21]. Additionally, our results match those of* Byz-EF21-BC *(a bidirectionally compressed method with error feedback for Byzantine-robust learning) from Rammal et al. [59, Theorem 3.1], in the absence of Byzantine workers.*

Proof [Proof of Theorem 25] Let $A_i, B_i > 0$ be some constants to be determined later, and define

$$\Psi^k := f(X^k) - f^\star + \sum_{i=1}^p A_i \frac{1}{n}\sum_{j=1}^n \left\|\nabla_i f_j(X^k) - G_{i,j}^k\right\|_{(i)\star}^2 + \sum_{i=1}^p B_i \left\|X_i^k - W_i^k\right\|_{(i)}^2 .$$

Step I: Bounding $\mathbb{E}\left[\left\|\nabla_i f_j(X^{k+1}) - G_{i,j}^{k+1}\right\|_{(i)\star}^2\right]$. The algorithm's update rule gives

$$\begin{aligned}&\mathbb{E}\left[\left\|\nabla_i f_j(X^{k+1}) - G_{i,j}^{k+1}\right\|_{(i)\star}^2 \middle| X^{k+1}, W^{k+1}, G_{i,j}^k\right] \\ &= \quad \mathbb{E}\left[\left\|\nabla_i f_j(X^{k+1}) - G_{i,j}^k - \mathcal{C}_{i,j}^k(\nabla_i f_j(W^{k+1}) - G_{i,j}^k)\right\|_{(i)\star}^2 \middle| X^{k+1}, W^{k+1}, G_{i,j}^k\right] \\ &\overset{(27)}{\leq} \quad \left(1 + \frac{\alpha_D}{2}\right)\mathbb{E}\left[\left\|\nabla_i f_j(W^{k+1}) - G_{i,j}^k - \mathcal{C}_{i,j}^k(\nabla_i f_j(W^{k+1}) - G_{i,j}^k)\right\|_{(i)\star}^2 \middle| X^{k+1}, W^{k+1}, G_{i,j}^k\right] \\ &\qquad + \left(1 + \frac{2}{\alpha_D}\right)\mathbb{E}\left[\left\|\nabla_i f_j(X^{k+1}) - \nabla_i f_j(W^{k+1})\right\|_{(i)\star}^2 \middle| X^{k+1}, W^{k+1}, G_{i,j}^k\right] \\ &\leq \quad \left(1 + \frac{\alpha_D}{2}\right)(1 - \alpha_D)\mathbb{E}\left[\left\|\nabla_i f_j(W^{k+1}) - G_{i,j}^k\right\|_{(i)\star}^2 \middle| X^{k+1}, W^{k+1}, G_{i,j}^k\right] \\ &\qquad + \left(1 + \frac{2}{\alpha_D}\right)\mathbb{E}\left[\left\|\nabla_i f_j(X^{k+1}) - \nabla_i f_j(W^{k+1})\right\|_{(i)\star}^2 \middle| X^{k+1}, W^{k+1}, G_{i,j}^k\right] \\ &\overset{(29)}{\leq} \quad \left(1 - \frac{\alpha_D}{2}\right)\left\|\nabla_i f_j(W^{k+1}) - G_{i,j}^k\right\|_{(i)\star}^2 + \left(1 + \frac{2}{\alpha_D}\right)\left\|\nabla_i f_j(X^{k+1}) - \nabla_i f_j(W^{k+1})\right\|_{(i)\star}^2 \\ &\overset{(27)}{\leq} \quad \left(1 - \frac{\alpha_D}{2}\right)\left(1 + \frac{\alpha_D}{4}\right)\left\|\nabla_i f_j(X^k) - G_{i,j}^k\right\|_{(i)\star}^2 \\ &\qquad + \left(1 - \frac{\alpha_D}{2}\right)\left(1 + \frac{4}{\alpha_D}\right)\left\|\nabla_i f_j(W^{k+1}) - \nabla_i f_j(X^k)\right\|_{(i)\star}^2\end{aligned}$$

$$\begin{aligned}
& + \left(1 + \frac{2}{\alpha_D}\right) \left\|\nabla_i f_j(X^{k+1}) - \nabla_i f_j(W^{k+1})\right\|_{(i)\star}^2 \\
\overset{(29),(30)}{\leq} \quad & \left(1 - \frac{\alpha_D}{4}\right) \left\|\nabla_i f_j(X^k) - G_{i,j}^k\right\|_{(i)\star}^2 + \frac{4}{\alpha_D} \left\|\nabla_i f_j(W^{k+1}) - \nabla_i f_j(X^k)\right\|_{(i)\star}^2 \\
& + \left(1 + \frac{2}{\alpha_D}\right) \left\|\nabla_i f_j(X^{k+1}) - \nabla_i f_j(W^{k+1})\right\|_{(i)\star}^2 .
\end{aligned}$$

Therefore, using smoothness,

$$\begin{aligned}
& \mathbb{E}\left[\left\|\nabla_i f_j(X^{k+1}) - G_{i,j}^{k+1}\right\|_{(i)\star}^2 \middle| X^{k+1}, W^{k+1}, G_{i,j}^k \right] \\
& \overset{(7)}{\leq} \quad \left(1 - \frac{\alpha_D}{4}\right) \left\|\nabla_i f_j(X^k) - G_{i,j}^k\right\|_{(i)\star}^2 + \frac{4}{\alpha_D} (L_{i,j}^0)^2 \left\|W_i^{k+1} - X_i^k\right\|_{(i)}^2 \\
& \qquad + \left(1 + \frac{2}{\alpha_D}\right) (L_{i,j}^0)^2 \left\|X_i^{k+1} - W_i^{k+1}\right\|_{(i)}^2 \\
& \overset{(27)}{\leq} \quad \left(1 - \frac{\alpha_D}{4}\right) \left\|\nabla_i f_j(X^k) - G_{i,j}^k\right\|_{(i)\star}^2 + \frac{8}{\alpha_D} (L_{i,j}^0)^2 \left\|X_i^{k+1} - X_i^k\right\|_{(i)}^2 \\
& \qquad + \frac{8}{\alpha_D} (L_{i,j}^0)^2 \left\|X_i^{k+1} - W_i^{k+1}\right\|_{(i)}^2 + \left(1 + \frac{2}{\alpha_D}\right) (L_{i,j}^0)^2 \left\|X_i^{k+1} - W_i^{k+1}\right\|_{(i)}^2 \\
& \leq \quad \left(1 - \frac{\alpha_D}{4}\right) \left\|\nabla_i f_j(X^k) - G_{i,j}^k\right\|_{(i)\star}^2 + \frac{8}{\alpha_D} (L_{i,j}^0)^2 \gamma_i^2 \left\|G_i^k\right\|_{(i)\star}^2 \\
& \qquad + \frac{11}{\alpha_D} (L_{i,j}^0)^2 \left\|X_i^{k+1} - W_i^{k+1}\right\|_{(i)}^2 .
\end{aligned}$$

Taking expectation, we obtain the recursion

$$\begin{aligned}
& \mathbb{E}\left[\left\|\nabla_i f_j(X^{k+1}) - G_{i,j}^{k+1}\right\|_{(i)\star}^2 \right] \\
& \quad \leq \quad \left(1 - \frac{\alpha_D}{4}\right) \mathbb{E}\left[\left\|\nabla_i f_j(X^k) - G_{i,j}^k\right\|_{(i)\star}^2 \right] + \frac{8}{\alpha_D} (L_{i,j}^0)^2 \gamma_i^2 \mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2 \right] \\
& \qquad\quad + \frac{11}{\alpha_D} (L_{i,j}^0)^2 \mathbb{E}\left[\left\|X_i^{k+1} - W_i^{k+1}\right\|_{(i)}^2 \right] .
\end{aligned} \tag{12}$$

Step II: Bounding $\mathbb{E}\left[\left\|X_i^{k+1} - W_i^{k+1}\right\|_{(i)}^2 \right]$. By Theorem 16

$$\mathbb{E}\left[\left\|X_i^{k+1} - W_i^{k+1}\right\|_{(i)}^2 \right] \leq \left(1 - \frac{\alpha_P}{2}\right) \mathbb{E}\left[\left\|X_i^k - W_i^k\right\|_{(i)}^2 \right] + \frac{2}{\alpha_P} \gamma_i^2 \mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2 \right] . \tag{13}$$

Step III: Bounding Ψ^{k+1}. By Theorem 14 and Jensen's inequality

$$\begin{aligned}
& \Psi^{k+1} \\
& = \quad f(X^{k+1}) - f^\star + \sum_{i=1}^p A_i \frac{1}{n} \sum_{j=1}^n \left\|\nabla_i f_j(X^{k+1}) - G_{i,j}^{k+1}\right\|_{(i)\star}^2 + \sum_{i=1}^p B_i \left\|X_i^{k+1} - W_i^{k+1}\right\|_{(i)}^2 \\
& \leq \quad f(X^k) - f^\star + \sum_{i=1}^p \frac{3\gamma_i}{2} \left\|\nabla_i f(X^k) - G_i^k\right\|_{(i)\star}^2 - \sum_{i=1}^p \frac{\gamma_i}{4} \left\|\nabla_i f(X^k)\right\|_{(i)\star}^2
\end{aligned}$$

$$
\begin{aligned}
& -\sum_{i=1}^{p}\left(\frac{1}{4\gamma_i}-\frac{L_i^0}{2}\right)\gamma_i^2\left\|G_i^k\right\|_{(i)\star}^2 \\
& +\sum_{i=1}^{p} A_i \frac{1}{n}\sum_{j=1}^{n}\left\|\nabla_i f_j(X^{k+1})-G_{i,j}^{k+1}\right\|_{(i)\star}^2+\sum_{i=1}^{p} B_i\left\|X_i^{k+1}-W_i^{k+1}\right\|_{(i)}^2 \\
\leq\ & f(X^k)-f^\star+\sum_{i=1}^{p}\frac{3\gamma_i}{2}\frac{1}{n}\sum_{j=1}^{n}\left\|\nabla_i f_j(X^k)-G_{i,j}^k\right\|_{(i)\star}^2-\sum_{i=1}^{p}\frac{\gamma_i}{4}\left\|\nabla_i f(X^k)\right\|_{(i)\star}^2 \\
& -\sum_{i=1}^{p}\left(\frac{1}{4\gamma_i}-\frac{L_i^0}{2}\right)\gamma_i^2\left\|G_i^k\right\|_{(i)\star}^2 \\
& +\sum_{i=1}^{p} A_i \frac{1}{n}\sum_{j=1}^{n}\left\|\nabla_i f_j(X^{k+1})-G_{i,j}^{k+1}\right\|_{(i)\star}^2+\sum_{i=1}^{p} B_i\left\|X_i^{k+1}-W_i^{k+1}\right\|_{(i)}^2 .
\end{aligned}
$$

Taking expectation and using (12) gives

$$
\begin{aligned}
& \mathbb{E}\left[\Psi^{k+1}\right] \\
& \leq \mathbb{E}\left[f(X^k)-f^\star\right]+\sum_{i=1}^{p}\frac{3\gamma_i}{2}\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|\nabla_i f_j(X^k)-G_{i,j}^k\right\|_{(i)\star}^2\right]-\sum_{i=1}^{p}\frac{\gamma_i}{4}\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}^2\right] \\
& \quad -\sum_{i=1}^{p}\left(\frac{1}{4\gamma_i}-\frac{L_i^0}{2}\right)\gamma_i^2\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right]+\sum_{i=1}^{p} A_i\frac{1}{n}\sum_{j=1}^{n}\left(1-\frac{\alpha_D}{4}\right)\mathbb{E}\left[\left\|\nabla_i f_j(X^k)-G_{i,j}^k\right\|_{(i)\star}^2\right] \\
& \quad +\sum_{i=1}^{p} A_i\frac{1}{n}\sum_{j=1}^{n}\frac{8}{\alpha_D}(L_{i,j}^0)^2\gamma_i^2\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right]+\sum_{i=1}^{p} A_i\frac{1}{n}\sum_{j=1}^{n}\frac{11}{\alpha_D}(L_{i,j}^0)^2\mathbb{E}\left[\left\|X_i^{k+1}-W_i^{k+1}\right\|_{(i)}^2\right] \\
& \quad +\sum_{i=1}^{p} B_i\mathbb{E}\left[\left\|X_i^{k+1}-W_i^{k+1}\right\|_{(i)}^2\right] \\
& = \mathbb{E}\left[f(X^k)-f^\star\right]+\sum_{i=1}^{p}\left(\frac{3\gamma_i}{2}+A_i\left(1-\frac{\alpha_D}{4}\right)\right)\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|\nabla_i f_j(X^k)-G_{i,j}^k\right\|_{(i)\star}^2\right] \\
& \quad -\sum_{i=1}^{p}\frac{\gamma_i}{4}\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}^2\right]-\sum_{i=1}^{p}\left(\frac{1}{4\gamma_i}-\frac{L_i^0}{2}-A_i\frac{8}{\alpha_D}(\tilde{L}_i^0)^2\right)\gamma_i^2\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right] \\
& \quad +\sum_{i=1}^{p}\left(A_i\frac{11}{\alpha_D}(\tilde{L}_i^0)^2+B_i\right)\mathbb{E}\left[\left\|X_i^{k+1}-W_i^{k+1}\right\|_{(i)}^2\right].
\end{aligned}
$$

Next, applying (13), we get

$$
\begin{aligned}
\mathbb{E}\left[\Psi^{k+1}\right] \leq\ & \mathbb{E}\left[f(X^k)-f^\star\right]+\sum_{i=1}^{p}\left(\frac{3\gamma_i}{2}+A_i\left(1-\frac{\alpha_D}{4}\right)\right)\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|\nabla_i f_j(X^k)-G_{i,j}^k\right\|_{(i)\star}^2\right] \\
& -\sum_{i=1}^{p}\frac{\gamma_i}{4}\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}^2\right]-\sum_{i=1}^{p}\left(\frac{1}{4\gamma_i}-\frac{L_i^0}{2}-A_i\frac{8}{\alpha_D}(\tilde{L}_i^0)^2\right)\gamma_i^2\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{i=1}^{p}\left(A_i\frac{11}{\alpha_D}(\tilde{L}_i^0)^2+B_i\right)\frac{2}{\alpha_P}\gamma_i^2\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right]\\
&+\sum_{i=1}^{p}\left(A_i\frac{11}{\alpha_D}(\tilde{L}_i^0)^2+B_i\right)\left(1-\frac{\alpha_P}{2}\right)\mathbb{E}\left[\left\|X_i^k-W_i^k\right\|_{(i)}^2\right].
\end{aligned}
$$

Taking $A_i=\frac{6\gamma_i}{\alpha_D}$ and $B_i=A_i\frac{11}{\alpha_D}\left(\frac{2}{\alpha_P}-1\right)(\tilde{L}_i^0)^2=\frac{66\gamma_i}{\alpha_D^2}\left(\frac{2}{\alpha_P}-1\right)(\tilde{L}_i^0)^2$ yields

$$
\begin{aligned}
\frac{3\gamma_i}{2}+A_i\left(1-\frac{\alpha_D}{4}\right)&=A_i,\\
\left(A_i\frac{11}{\alpha_D}(\tilde{L}_i^0)^2+B_i\right)\left(1-\frac{\alpha_P}{2}\right)&=B_i,
\end{aligned}
$$

and consequently,

$$
\begin{aligned}
\mathbb{E}\left[\Psi^{k+1}\right] \;\leq\;& \mathbb{E}\left[f(X^k)-f^\star\right]+\sum_{i=1}^{p}A_i\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|\nabla_i f_j(X^k)-G_{i,j}^k\right\|_{(i)\star}^2\right]\\
&-\sum_{i=1}^{p}\frac{\gamma_i}{4}\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}^2\right]-\sum_{i=1}^{p}\left(\frac{1}{4\gamma_i}-\frac{L_i^0}{2}-\frac{8A_i}{\alpha_D}(\tilde{L}_i^0)^2\right)\gamma_i^2\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right]\\
&+\sum_{i=1}^{p}\frac{B_i}{1-\frac{\alpha_P}{2}}\frac{2}{\alpha_P}\gamma_i^2\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right]+\sum_{i=1}^{p}B_i\mathbb{E}\left[\left\|X_i^k-W_i^k\right\|_{(i)}^2\right]\\
=\;& \mathbb{E}\left[\Psi^k\right]-\sum_{i=1}^{p}\frac{\gamma_i}{4}\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}^2\right]\\
&-\sum_{i=1}^{p}\left(\frac{1}{4\gamma_i}-\frac{L_i^0}{2}-\frac{8A_i}{\alpha_D}(\tilde{L}_i^0)^2-\frac{4B_i}{\alpha_P(2-\alpha_P)}\right)\gamma_i^2\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right].
\end{aligned}
$$

Now, note that

$$
\frac{1}{4\gamma_i}-\frac{L_i^0}{2}-\frac{8A_i}{\alpha_D}(\tilde{L}_i^0)^2-\frac{4B_i}{\alpha_P(2-\alpha_P)}=\frac{1}{4\gamma_i}-\frac{L_i^0}{2}-\underbrace{\left(\frac{48}{\alpha_D^2}(\tilde{L}_i^0)^2+\frac{264}{\alpha_P^2\alpha_D^2}(\tilde{L}_i^0)^2\right)}_{:=\zeta_i}\gamma_i\geq 0
$$

for $\gamma_i\leq\frac{1}{2L_i^0+2\sqrt{\zeta_i}}$. For such a choice of the stepsizes, we have

$$
\mathbb{E}\left[\Psi^{k+1}\right]\leq\mathbb{E}\left[\Psi^k\right]-\sum_{i=1}^{p}\frac{\gamma_i}{4}\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}^2\right],
$$

and hence

$$
\sum_{k=0}^{K-1}\sum_{i=1}^{p}\gamma_i\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}^2\right]\leq 4\sum_{k=0}^{K-1}\left(\mathbb{E}\left[\Psi^k\right]-\mathbb{E}\left[\Psi^{k+1}\right]\right)\leq 4\Psi^0.
$$

Lastly, dividing by $\frac{K}{p}\sum_{l=1}^{p}\gamma_l$, we obtain

$$\frac{1}{K}\sum_{k=0}^{K-1}\sum_{i=1}^{p}\frac{\gamma_i}{\frac{1}{p}\sum_{l=1}^{p}\gamma_l}\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}^2\right] \leq \frac{1}{K}\frac{4\Psi^0}{\frac{1}{p}\sum_{l=1}^{p}\gamma_l}.$$

■

F.3.2. LAYER-WISE (L^0, L^1)–SMOOTH REGIME

We now consider a deterministic variant of EF21-Muon (Algorithm 2) without primal compression, which iterates

$$\begin{aligned} X_i^{k+1} &= \mathrm{LMO}_{\mathcal{B}(X_i^k, t_i^k)}\left(G_i^k\right), \\ G_{i,j}^{k+1} &= G_{i,j}^k + \mathcal{C}_{i,j}^k(\nabla_i f_j(X^{k+1}) - G_{i,j}^k), \\ G_i^{k+1} &= \frac{1}{n}\sum_{j=1}^{n} G_{i,j}^{k+1} = G_i^k + \frac{1}{n}\sum_{j=1}^{n}\mathcal{C}_{i,j}^k(\nabla_i f_j(X^{k+1}) - G_{i,j}^k). \end{aligned}$$

This corresponds to using identity compressors on the server side.

Theorem 28 *Let Assumptions 1, 2, 8 and 9 hold and let $\{X^k\}_{k=0}^{K-1}$, $K \geq 1$, be the iterates of Algorithm 2 run with $\mathcal{C}_i^k \equiv \mathcal{I}$ (the identity compressor), $\mathcal{C}_{i,j}^k \in \mathbb{B}_\star(\alpha_D)$, and*

$$t_i^k \equiv t_i = \frac{\eta_i}{\sqrt{K+1}}, \qquad i = 1, \ldots, p,$$

for some $\eta_i > 0$. Then,

$$\begin{aligned} &\min_{k=0,\ldots,K}\sum_{i=1}^{p}\frac{\eta_i}{\frac{1}{p}\sum_{l=1}^{p}\eta_i}\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right] \\ \leq\quad & \frac{\exp\left(4\max_{i\in[p],j\in[n]}(\eta_i^2 C_i L_{i,j}^1)\right)}{\sqrt{K+1}\left(\frac{1}{p}\sum_{l=1}^{p}\eta_i\right)}\Psi^0 \\ & + \frac{1}{\sqrt{K+1}\left(\frac{1}{p}\sum_{l=1}^{p}\eta_i\right)}\left(\frac{1}{n}\sum_{j=1}^{n}4\max_{i\in[p]}(\eta_i^2 C_i L_{i,j}^1)\left(f^\star - f_j^\star\right) + \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{p}\frac{\eta_i^2 C_i L_{i,j}^0}{L_{i,j}^1} + \sum_{i=1}^{p}\eta_i^2 D_i\right). \end{aligned}$$

where $C_i := \frac{L_i^1}{2} + \frac{2\sqrt{1-\alpha_D}L_{i,\max}^1}{1-\sqrt{1-\alpha_D}}$, $D_i := \frac{L_i^0}{2} + \frac{2\sqrt{1-\alpha_D}\bar{L}_i^0}{1-\sqrt{1-\alpha_D}}$ and

$$\Psi^k := f(X^k) - f^\star + \sum_{i=1}^{p}\frac{2t_i}{1-\sqrt{1-\alpha_D}}\frac{1}{n}\sum_{j=1}^{n}\left\|\nabla_i f_j(X^k) - G_{i,j}^k\right\|_{(i)\star}.$$

Remark 29 *Theorem 4 follows as a corollary of the result in Theorem 28 by setting $p = 1$ and initializing with $G_j^0 = \nabla f_j(X^0)$ for all $j \in [n]$.*

Proof Let $A_i > 0$ be some constants to be determined later, and define

$$\Psi^k := f(X^k) - f^\star + \sum_{i=1}^p A_i \frac{1}{n} \sum_{j=1}^n \left\| \nabla_i f_j(X^k) - G_{i,j}^k \right\|_{(i)\star}.$$

By Theorem 15 and Jensen's inequality

$$\begin{aligned}
\Psi^{k+1} &= f(X^{k+1}) - f^\star + \sum_{i=1}^p A_i \frac{1}{n} \sum_{j=1}^n \left\| \nabla_i f_j(X^{k+1}) - G_{i,j}^{k+1} \right\|_{(i)\star} \\
&\leq f(X^k) - f^\star + \sum_{i=1}^p 2t_i \left\| \nabla_i f(X^k) - G_i^k \right\|_{(i)\star} - \sum_{i=1}^p t_i \left\| \nabla_i f(X^k) \right\|_{(i)\star} \\
&\quad + \sum_{i=1}^p \frac{L_i^0 + L_i^1 \left\| \nabla_i f(X^k) \right\|_{(i)\star}}{2} t_i^2 + \sum_{i=1}^p A_i \frac{1}{n} \sum_{j=1}^n \left\| \nabla_i f_j(X^{k+1}) - G_{i,j}^{k+1} \right\|_{(i)\star} \\
&\leq f(X^k) - f^\star + \sum_{i=1}^p 2t_i \frac{1}{n} \sum_{j=1}^n \left\| \nabla_i f_j(X^k) - G_{i,j}^k \right\|_{(i)\star} - \sum_{i=1}^p t_i \left\| \nabla_i f(X^k) \right\|_{(i)\star} \\
&\quad + \sum_{i=1}^p \frac{L_i^0 + L_i^1 \left\| \nabla_i f(X^k) \right\|_{(i)\star}}{2} t_i^2 + \sum_{i=1}^p A_i \frac{1}{n} \sum_{j=1}^n \left\| \nabla_i f_j(X^{k+1}) - G_{i,j}^{k+1} \right\|_{(i)\star}.
\end{aligned}$$

Taking expectation conditioned on $[X^{k+1}, X^k, G^k]$ and using Theorem 19 gives

$$\begin{aligned}
&\mathbb{E}\left[\Psi^{k+1} \middle| X^{k+1}, X^k, G^k \right] \\
\leq\ & f(X^k) - f^\star + \sum_{i=1}^p 2t_i \frac{1}{n} \sum_{j=1}^n \left\| \nabla_i f_j(X^k) - G_{i,j}^k \right\|_{(i)\star} - \sum_{i=1}^p t_i \left\| \nabla_i f(X^k) \right\|_{(i)\star} \\
& + \sum_{i=1}^p \frac{L_i^0 + L_i^1 \left\| \nabla_i f(X^k) \right\|_{(i)\star}}{2} t_i^2 \\
& + \sum_{i=1}^p A_i \frac{1}{n} \sum_{j=1}^n \mathbb{E}\left[\left\| \nabla_i f_j(X^{k+1}) - G_{i,j}^{k+1} \right\|_{(i)\star} \middle| X^{k+1}, X^k, G^k \right] \\
\overset{(19)}{\leq}\ & f(X^k) - f^\star + \sum_{i=1}^p 2t_i \frac{1}{n} \sum_{j=1}^n \left\| \nabla_i f_j(X^k) - G_{i,j}^k \right\|_{(i)\star} - \sum_{i=1}^p t_i \left\| \nabla_i f(X^k) \right\|_{(i)\star} \\
& + \sum_{i=1}^p \frac{L_i^0 + L_i^1 \left\| \nabla_i f(X^k) \right\|_{(i)\star}}{2} t_i^2 \\
& + \sum_{i=1}^p A_i \sqrt{1-\alpha_D} \frac{1}{n} \sum_{j=1}^n \left(\left\| \nabla_i f_j(X^k) - G_{i,j}^k \right\|_{(i)\star} + \left(L_{i,j}^0 + L_{i,j}^1 \left\| \nabla_i f_j(X^k) \right\|_{(i)\star} \right) t_i \right) \\
=\ & f(X^k) - f^\star + \sum_{i=1}^p \left(2t_i + A_i \sqrt{1-\alpha_D} \right) \frac{1}{n} \sum_{j=1}^n \left\| \nabla_i f_j(X^k) - G_{i,j}^k \right\|_{(i)\star} - \sum_{i=1}^p t_i \left\| \nabla_i f(X^k) \right\|_{(i)\star}
\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{i=1}^{p} \frac{L_i^1 t_i^2}{2} \left\| \nabla_i f(X^k) \right\|_{(i)\star} + \sqrt{1-\alpha_D} \sum_{i=1}^{p} A_i t_i \left(\frac{1}{n} \sum_{j=1}^{n} L_{i,j}^1 \left\| \nabla_i f_j(X^k) \right\|_{(i)\star} \right) \\
&+ \sum_{i=1}^{p} \frac{t_i^2 L_i^0}{2} + \sqrt{1-\alpha_D} \sum_{i=1}^{p} A_i t_i \bar{L}_i^0.
\end{aligned}
$$

Now, letting $A_i = \frac{2t_i}{1-\sqrt{1-\alpha_D}}$, we have

$$
2t_i + A_i \sqrt{1-\alpha_D} = 2t_i + \frac{2t_i}{1-\sqrt{1-\alpha_D}} \sqrt{1-\alpha_D} = A_i,
$$

and consequently,

$$
\begin{aligned}
&\mathbb{E}\left[\Psi^{k+1} \middle| X^{k+1}, X^k, G^k \right] \\
\leq\ & f(X^k) - f^\star + \sum_{i=1}^{p} A_i \frac{1}{n} \sum_{j=1}^{n} \left\| \nabla_i f_j(X^k) - G_{i,j}^k \right\|_{(i)\star} - \sum_{i=1}^{p} t_i \left\| \nabla_i f(X^k) \right\|_{(i)\star} \\
&+ \sum_{i=1}^{p} \frac{L_i^1 t_i^2}{2} \left\| \nabla_i f(X^k) \right\|_{(i)\star} + \sqrt{1-\alpha_D} \sum_{i=1}^{p} A_i t_i \left(\frac{1}{n} \sum_{j=1}^{n} L_{i,j}^1 \left\| \nabla_i f_j(X^k) \right\|_{(i)\star} \right) \\
&+ \sum_{i=1}^{p} \frac{t_i^2 L_i^0}{2} + \sqrt{1-\alpha_D} \sum_{i=1}^{p} A_i t_i \bar{L}_i^0 \\
\leq\ & \Psi^k - \sum_{i=1}^{p} t_i \left\| \nabla_i f(X^k) \right\|_{(i)\star} + \sum_{i=1}^{p} \frac{L_i^1 t_i^2}{2} \left(\frac{1}{n} \sum_{j=1}^{n} \left\| \nabla_i f_j(X^k) \right\|_{(i)\star} \right) \\
&+ \sum_{i=1}^{p} \frac{2\sqrt{1-\alpha_D} L_{i,\max}^1}{1-\sqrt{1-\alpha_D}} t_i^2 \left(\frac{1}{n} \sum_{j=1}^{n} \left\| \nabla_i f_j(X^k) \right\|_{(i)\star} \right) + \sum_{i=1}^{p} \left(\frac{t_i^2 L_i^0}{2} + \frac{2\sqrt{1-\alpha_D}}{1-\sqrt{1-\alpha_D}} t_i^2 \bar{L}_i^0 \right) \\
=\ & \Psi^k - \sum_{i=1}^{p} t_i \left\| \nabla_i f(X^k) \right\|_{(i)\star} \\
&+ \sum_{i=1}^{p} \left(\underbrace{\left(\frac{L_i^1}{2} + \frac{2\sqrt{1-\alpha_D} L_{i,\max}^1}{1-\sqrt{1-\alpha_D}} \right)}_{:=C_i} \frac{1}{n} \sum_{j=1}^{n} \left\| \nabla_i f_j(X^k) \right\|_{(i)\star} + \underbrace{\frac{L_i^0}{2} + \frac{2\sqrt{1-\alpha_D} \bar{L}_i^0}{1-\sqrt{1-\alpha_D}}}_{:=D_i} \right) t_i^2.
\end{aligned}
$$

Taking $t_i = \frac{\eta_i}{\sqrt{K+1}}$ for some $\eta_i > 0$ and using Theorem 24 with $x_i = \eta_i^2 C_i$, we get

$$
\begin{aligned}
&\mathbb{E}\left[\Psi^{k+1} \middle| X^{k+1}, X^k, G^k \right] \\
\leq\ & \Psi^k - \sum_{i=1}^{p} t_i \left\| \nabla_i f(X^k) \right\|_{(i)\star} + \frac{1}{K+1} \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{p} \eta_i^2 C_i \left\| \nabla_i f_j(X^k) \right\|_{(i)\star} + \sum_{i=1}^{p} D_i t_i^2 \\
\overset{(24)}{\leq}\ & \Psi^k - \sum_{i=1}^{p} t_i \left\| \nabla_i f(X^k) \right\|_{(i)\star} + \sum_{i=1}^{p} D_i t_i^2 + \frac{1}{K+1} \frac{1}{n} \sum_{j=1}^{n} 4 \max_{i \in [p]} (\eta_i^2 C_i L_{i,j}^1) \left(f_j(X^k) - f^\star \right)
\end{aligned}
$$

$$\begin{aligned}
&+\frac{1}{K+1}\frac{1}{n}\sum_{j=1}^{n}4\max_{i\in[p]}(\eta_i^2 C_i L^1_{i,j})\left(f^\star - f_j^\star\right)+\frac{1}{K+1}\frac{1}{n}\sum_{j=1}^{n}\frac{\sum_{i=1}^{p}\eta_i^4 C_i^2 L^0_{i,j}}{\max_{i\in[p]}(\eta_i^2 C_i L^1_{i,j})}\\
\leq\quad & \Psi^k - \frac{1}{\sqrt{K+1}}\sum_{i=1}^{p}\eta_i\left\|\nabla_i f(X^k)\right\|_{(i)\star}+\frac{4}{K+1}\max_{i\in[p],j\in[n]}(\eta_i^2 C_i L^1_{i,j})\frac{1}{n}\sum_{j=1}^{n}\left(f_j(X^k)-f^\star\right)\\
&+\frac{1}{K+1}\left(\frac{1}{n}\sum_{j=1}^{n}4\max_{i\in[p]}(\eta_i^2 C_i L^1_{i,j})\left(f^\star - f_j^\star\right)+\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{p}\frac{\eta_i^2 C_i L^0_{i,j}}{L^1_{i,j}}+\sum_{i=1}^{p}\eta_i^2 D_i\right).
\end{aligned}$$

Now, since $\frac{1}{n}\sum_{j=1}^{n}\left(f_j(X^k)-f^\star\right)=f(X^k)-f^\star\leq\Psi^k$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\Psi^{k+1}\middle| X^{k+1},X^k,G^k\right]\\
\leq\quad & \left(1+\frac{4}{K+1}\max_{i\in[p],j\in[n]}(\eta_i^2 C_i L^1_{i,j})\right)\Psi^k-\frac{1}{\sqrt{K+1}}\sum_{i=1}^{p}\eta_i\left\|\nabla_i f(X^k)\right\|_{(i)\star}\\
&+\frac{1}{K+1}\left(\frac{1}{n}\sum_{j=1}^{n}4\max_{i\in[p]}(\eta_i^2 C_i L^1_{i,j})\left(f^\star - f_j^\star\right)+\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{p}\frac{\eta_i^2 C_i L^0_{i,j}}{L^1_{i,j}}+\sum_{i=1}^{p}\eta_i^2 D_i\right).
\end{aligned}$$

Taking expectation,

$$\begin{aligned}
&\mathbb{E}\left[\Psi^{k+1}\right]\\
\leq\quad & \left(1+\underbrace{\frac{4}{K+1}\max_{i\in[p],j\in[n]}(\eta_i^2 C_i L^1_{i,j})}_{:=a_1}\right)\mathbb{E}\left[\Psi^k\right]-\frac{1}{\sqrt{K+1}}\sum_{i=1}^{p}\eta_i\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right]\\
&+\underbrace{\frac{1}{K+1}\left(\frac{1}{n}\sum_{j=1}^{n}4\max_{i\in[p]}(\eta_i^2 C_i L^1_{i,j})\left(f^\star - f_j^\star\right)+\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{p}\frac{\eta_i^2 C_i L^0_{i,j}}{L^1_{i,j}}+\sum_{i=1}^{p}\eta_i^2 D_i\right)}_{:=a_2},
\end{aligned}$$

and hence, applying Theorem 42 with $A^k=\mathbb{E}\left[\Psi^k\right]$ and $B_i^k=\frac{\eta_i}{\sqrt{K+1}}\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right]$,

$$\begin{aligned}
&\min_{k=0,\ldots,K}\sum_{i=1}^{p}\frac{\eta_i}{\sqrt{K+1}}\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right]\\
\leq\quad & \frac{\exp\left(\frac{4}{K+1}\max_{i\in[p],j\in[n]}(\eta_i^2 C_i L^1_{i,j})(K+1)\right)}{(K+1)}\Psi^0\\
&+\frac{1}{K+1}\left(\frac{1}{n}\sum_{j=1}^{n}4\max_{i\in[p]}(\eta_i^2 C_i L^1_{i,j})\left(f^\star - f_j^\star\right)+\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{p}\frac{\eta_i^2 C_i L^0_{i,j}}{L^1_{i,j}}+\sum_{i=1}^{p}\eta_i^2 D_i\right).
\end{aligned}$$

Dividing by $\frac{\frac{1}{p}\sum_{l=1}^{p}\eta_i}{\sqrt{K+1}}$ finishes the proof. ■

F.4. Stochastic Setting

F.4.1. Layer-wise Smooth Regime

Theorem 30 *Let Assumptions 1, 6, 7 and 10 hold. Let $\{X^k\}_{k=0}^{K-1}$, $K \geq 1$, be the iterates of Algorithm 3 run with $\mathcal{C}_i^k \in \mathbb{B}(\alpha_P)$, $\mathcal{C}_{i,j}^k \in \mathbb{B}_2(\alpha_D)$, any $\beta_i \in (0,1]$, and*

$$0 \leq \gamma_i^k \equiv \gamma_i \leq \frac{1}{2L_i^0 + 2\sqrt{\zeta_i}}, \qquad i = 1, \ldots, p,$$

where $\zeta_i := \frac{\bar{\rho}_i^2}{\underline{\rho}_i^2}\left(\frac{12}{\beta_i^2}(L_i^0)^2 + \frac{24(\beta_i+2)}{\alpha_P^2}(L_i^0)^2 + \frac{36\left(\beta_i^2+4\right)}{\alpha_D^2}(\tilde{L}_i^0)^2 + \frac{144\beta_i^2(2\beta_i+5)}{\alpha_P^2\alpha_D^2}(\tilde{L}_i^0)^2\right)$. Then

$$\begin{aligned} &\frac{1}{K}\sum_{k=0}^{K-1}\sum_{i=1}^{p}\frac{\gamma_i}{\frac{1}{p}\sum_{l=1}^{p}\gamma_l}\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}^2\right] \\ \leq\ & \frac{1}{K}\frac{4\Psi^0}{\frac{1}{p}\sum_{l=1}^{p}\gamma_l} + 24\sum_{i=1}^{p}\left(\frac{1}{n} + \frac{(1-\alpha_D)\beta_i}{\alpha_D} + \frac{12\beta_i^2}{\alpha_D^2}\right)\frac{\sigma_i^2\bar{\rho}_i^2\beta_i\gamma_i}{\frac{1}{p}\sum_{l=1}^{p}\gamma_l}, \end{aligned} \tag{14}$$

where

$$\begin{aligned} \Psi^0 :=\ & f(X^0) - f^\star + \sum_{i=1}^{p}\frac{6\bar{\rho}_i^2}{\beta_i}\gamma_i\mathbb{E}\left[\left\|\nabla_i f(X^0) - M_i^0\right\|_2^2\right] \\ & + \sum_{i=1}^{p}\frac{72\bar{\rho}_i^2\beta_i}{\alpha_D^2}\gamma_i\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|\nabla_i f_j(X^0) - M_{i,j}^0\right\|_2^2\right] + \sum_{i=1}^{p}\frac{6\bar{\rho}_i^2}{\alpha_D}\gamma_i\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|M_{i,j}^0 - G_{i,j}^0\right\|_2^2\right] \end{aligned}$$

and $M_i^k := \frac{1}{n}\sum_{j=1}^{n} M_{i,j}^k$.

Corollary 31 *Let the assumptions of Theorem 30 hold and let $\{X^k\}_{k=0}^{K-1}$, $K \geq 1$, be the iterates of Algorithm 1 (Algorithm 3 with $p = 1$) run with $\mathcal{C}_i^k \in \mathbb{B}(\alpha_P)$, $\mathcal{C}_{i,j}^k \in \mathbb{B}_2(\alpha_D)$. Choosing the stepsize*

$$\gamma_1 = \frac{1}{\sqrt{2\zeta_1} + 2L_1^0} = \mathcal{O}\left(\left(\frac{\bar{\rho}_1^2 L_1^0}{\underline{\rho}_1^2\beta_1} + \frac{\bar{\rho}_1^2\tilde{L}_1^0}{\underline{\rho}_1^2\alpha_P\alpha_D}\right)^{-1}\right) \tag{15}$$

and momentum

$$\beta_1 = \min\left\{1, \left(\frac{\Psi^0 L_1^0 n}{\underline{\rho}_1^2\sigma_1^2 K}\right)^{1/2}, \left(\frac{\Psi^0 L_1^0\alpha_D}{\underline{\rho}_1^2\sigma_1^2 K}\right)^{1/3}, \left(\frac{\Psi^0 L_1^0\alpha_D^2}{\underline{\rho}_1^2\sigma_1^2 K}\right)^{1/4}\right\}, \tag{16}$$

the result in Theorem 30 guarantees that

$$\begin{aligned} &\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(X^k)\right\|_{\star}^2\right] \\ &\leq \mathcal{O}\left(\frac{\Psi^0\bar{\rho}_1^2\tilde{L}_1^0}{\underline{\rho}_1^2\alpha_P\alpha_D K} + \left(\frac{\Psi^0\bar{\rho}_1^4\sigma_1^2 L_1^0}{\underline{\rho}_1^2 nK}\right)^{1/2} + \left(\frac{\Psi^0\bar{\rho}_1^3\sigma_1 L_1^0}{\underline{\rho}_1^2\sqrt{\alpha_D}K}\right)^{2/3} + \left(\frac{\Psi^0\underline{\rho}_1^{8/3}\sigma_1^{2/3}L_1^0}{\bar{\rho}_1^2\alpha_D^{2/3}K}\right)^{3/4}\right). \end{aligned}$$

Remark 32 *Theorem 5 follows as a corollary of the result above by setting $p = 1$.*

Remark 33 *In the Euclidean case ($\bar{\rho}_i^2 = \underline{\rho}_i^2 = 1$), without primal compression ($\alpha_P = 1$), and for $p = 1$, the result in Theorem 30 simplifies to*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(X^k)\right\|^2\right] \leq \mathcal{O}\left(\frac{\Psi^0}{K\gamma} + \left(\frac{1}{n} + \frac{\beta}{\alpha_D} + \frac{\beta^2}{\alpha_D^2}\right)\beta\sigma^2\right),$$

for $\gamma \leq \mathcal{O}\left(\frac{\beta}{L_1^0} + \frac{\alpha_D}{\tilde{L}_1^0}\right)$, which recovers the rate of EF21-SDGM *in Fatkhullin et al. [13, Theorem 3] (up to a constant).*

Remark 34 *In the absence of stochasticity and momentum, i.e., when $\sigma_i^2 = 0$ and $\beta_i = 1$, and under the initialization $W^0 = X^0$, $M_j^0 = G_j^0 = \nabla f_j(X^0)$, Algorithm 3 reduces to Algorithm 2. In this setting, Theorem 30 guarantees that*

$$\frac{1}{K}\sum_{k=0}^{K-1}\sum_{i=1}^{p}\frac{\gamma_i}{\frac{1}{p}\sum_{l=1}^{p}\gamma_l}\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|^2_{(i)\star}\right] \leq \frac{1}{K}\frac{4\left(f(X^0) - f^\star\right)}{\frac{1}{p}\sum_{l=1}^{p}\gamma_l},$$

for

$$0 \leq \gamma_i^k \equiv \gamma_i \leq \frac{1}{2L_i^0 + 2\sqrt{\zeta_i}}, \qquad i = 1, \ldots, p,$$

where $\zeta_i := \frac{\bar{\rho}_i^2}{\underline{\rho}_i^2}\left(12(L_i^0)^2 + \frac{72}{\alpha_P^2}(L_i^0)^2 + \frac{180}{\alpha_D^2}(\tilde{L}_i^0)^2 + \frac{1008}{\alpha_P^2\alpha_D^2}(\tilde{L}_i^0)^2\right)$. This recovers the guarantee in Theorem 25, up to a constant factor.

Remark 35 *Alternatively, one may use compressors $\mathcal{C}_i^k \in \mathbb{B}_2(\alpha_P)$ in Theorem 30. The proof is essentially the same, with the only modification being the replacement of Theorem 16 by the recursion*

$$\begin{aligned}
&\mathbb{E}_{\mathcal{C}}\left[\left\|X_i^{k+1} - W_i^{k+1}\right\|_2^2\right] \\
=\quad & \mathbb{E}_{\mathcal{C}}\left[\left\|W_i^k + \mathcal{C}_i^k(X_i^{k+1} - W_i^k) - X_i^{k+1}\right\|_2^2\right] \\
\leq\quad & (1-\alpha_P)\left\|X_i^{k+1} - W_i^k\right\|_2^2 \\
\overset{(27)}{\leq}\quad & (1-\alpha_P)\left(1 + \frac{\alpha_P}{2}\right)\left\|X_i^k - W_i^k\right\|_2^2 + (1-\alpha_P)\left(1 + \frac{2}{\alpha_P}\right)\left\|X_i^{k+1} - X_i^k\right\|_2^2 \\
\overset{(29),(30)}{\leq}\quad & \left(1 - \frac{\alpha_P}{2}\right)\left\|X_i^k - W_i^k\right\|_2^2 + \frac{2\bar{\rho}_i^2}{\alpha_P}\left\|X_i^{k+1} - X_i^k\right\|_{(i)}^2 \\
=\quad & \left(1 - \frac{\alpha_P}{2}\right)\left\|X_i^k - W_i^k\right\|_2^2 + \frac{2\bar{\rho}_i^2}{\alpha_P}(\gamma_i^k)^2\left\|G_i^k\right\|_{(i)\star}^2.
\end{aligned}$$

The resulting convergence guarantee matches that of Theorem 30 up to a modification of the constant ζ_i, which now becomes

$$\zeta_i = \frac{\bar{\rho}_i^2}{\underline{\rho}_i^2}\left(\frac{12}{\beta_i^2}(L_i^0)^2 + \frac{24\bar{\rho}_i^2\left(\beta_i + 2\right)}{\alpha_P^2}(L_i^0)^2 + \frac{36\left(\beta_i^2 + 4\right)}{\alpha_D^2}(\tilde{L}_i^0)^2 + \frac{144\bar{\rho}_i^2\beta_i^2\left(2\beta_i + 5\right)}{\alpha_P^2\alpha_D^2}(\tilde{L}_i^0)^2\right),$$

where the additional norm equivalence factors highlighted in red arise due to the use of Euclidean compressors.

Proof [Proof of Theorem 30] Theorem 14 and Young's and Jensen's inequalities give

$$\begin{aligned} f(X^{k+1}) &\overset{(14)}{\leq} f(X^k) + \frac{3}{2}\sum_{i=1}^p \gamma_i \left\| \nabla_i f(X^k) - G_i^k \right\|_{(i)\star}^2 - \frac{1}{4}\sum_{i=1}^p \gamma_i \left\| \nabla_i f(X^k) \right\|_{(i)\star}^2 \\ &\quad - \sum_{i=1}^p \left(\frac{1}{4\gamma_i} - \frac{L_i^0}{2} \right) \gamma_i^2 \left\| G_i^k \right\|_{(i)\star}^2 \\ &\overset{(27)}{\leq} f(X^k) + 3\sum_{i=1}^p \bar{\rho}_i^2 \gamma_i \left(\left\| \nabla_i f(X^k) - M_i^k \right\|_2^2 + \frac{1}{n}\sum_{j=1}^n \left\| M_{i,j}^k - G_{i,j}^k \right\|_2^2 \right) \\ &\quad - \frac{1}{4}\sum_{i=1}^p \gamma_i \left\| \nabla_i f(X^k) \right\|_{(i)\star}^2 - \sum_{i=1}^p \left(\frac{1}{4\gamma_i} - \frac{L_i^0}{2} \right) \gamma_i^2 \left\| G_i^k \right\|_{(i)\star}^2 . \end{aligned}$$

Recall that by Lemmas 16, 17 and 18, we have

$$\begin{aligned} \mathbb{E}\left[\left\| X_i^{k+1} - W_i^{k+1} \right\|_{(i)}^2 \right] &\overset{(16)}{\leq} \left(1 - \frac{\alpha_P}{2}\right) \mathbb{E}\left[\left\| X_i^k - W_i^k \right\|_{(i)}^2 \right] + \frac{2}{\alpha_P}\gamma_i^2 \mathbb{E}\left[\left\| G_i^k \right\|_{(i)\star}^2 \right], \\ \mathbb{E}\left[\left\| M_{i,j}^{k+1} - G_{i,j}^{k+1} \right\|_2^2 \right] &\overset{(17)}{\leq} \left(1 - \frac{\alpha_D}{2}\right) \mathbb{E}\left[\left\| M_{i,j}^k - G_{i,j}^k \right\|_2^2 \right] + \frac{6\beta_i^2}{\alpha_D} \mathbb{E}\left[\left\| M_{i,j}^k - \nabla_i f_j(X^k) \right\|_2^2 \right] \\ &\quad + \frac{6\beta_i^2 (L_{i,j}^0)^2}{\alpha_D \underline{\rho}_i^2} \gamma_i^2 \mathbb{E}\left[\left\| G_i^k \right\|_\star^2 \right] + \frac{6\beta_i^2 (L_{i,j}^0)^2}{\alpha_D \underline{\rho}_i^2} \mathbb{E}\left[\left\| X_i^{k+1} - W_i^{k+1} \right\|_{(i)}^2 \right] \\ &\quad + (1 - \alpha_D)\beta_i^2 \sigma_i^2, \\ \mathbb{E}\left[\left\| \nabla_i f_j(X^{k+1}) - M_{i,j}^{k+1} \right\|_2^2 \right] &\overset{(18)}{\leq} \left(1 - \frac{\beta_i}{2}\right) \mathbb{E}\left[\left\| \nabla_i f_j(X^k) - M_{i,j}^k \right\|_2^2 \right] + \frac{2(L_{i,j}^0)^2}{\beta_i \underline{\rho}_i^2} \gamma_i^2 \mathbb{E}\left[\left\| G_i^k \right\|_{(i)\star}^2 \right] \\ &\quad + \frac{\beta_i^2}{\underline{\rho}_i^2}\left(1 + \frac{2}{\beta_i}\right)(L_{i,j}^0)^2 \mathbb{E}\left[\left\| X_i^{k+1} - W_i^{k+1} \right\|_{(i)}^2 \right] + \beta_i^2 \sigma_i^2, \\ \mathbb{E}\left[\left\| \nabla_i f(X^{k+1}) - M_i^{k+1} \right\|_2^2 \right] &\overset{(18)}{\leq} \left(1 - \frac{\beta_i}{2}\right) \mathbb{E}\left[\left\| \nabla_i f(X^k) - M_i^k \right\|_2^2 \right] + \frac{2(L_i^0)^2}{\beta_i \underline{\rho}_i^2} \gamma_i^2 \mathbb{E}\left[\left\| G_i^k \right\|_{(i)\star}^2 \right] \\ &\quad + \frac{\beta_i^2}{\underline{\rho}_i^2}\left(1 + \frac{2}{\beta_i}\right)(L_i^0)^2 \mathbb{E}\left[\left\| X_i^{k+1} - W_i^{k+1} \right\|_{(i)}^2 \right] + \frac{\beta_i^2 \sigma_i^2}{n}, \end{aligned}$$

where $M_i^k := \frac{1}{n}\sum_{j=1}^n M_{i,j}^k$. To simplify the notation, let us define $\delta^k := \mathbb{E}\left[f(X^k) - f^\star \right]$, $P_i^k := \gamma_i \mathbb{E}\left[\left\| \nabla_i f(X^k) - M_i^k \right\|_2^2 \right]$, $\tilde{P}_i^k := \gamma_i \frac{1}{n}\sum_{j=1}^n \mathbb{E}\left[\left\| \nabla_i f_j(X^k) - M_{i,j}^k \right\|_2^2 \right]$, $\tilde{S}_i^k := \gamma_i \frac{1}{n}\sum_{j=1}^n \mathbb{E}\left[\left\| M_{i,j}^k - G_{i,j}^k \right\|_2^2 \right]$ and $R_i^k := \gamma_i \mathbb{E}\left[\left\| X_i^k - W_i^k \right\|_{(i)}^2 \right]$. Then, the above inequalities yield

$$\delta^{k+1} \leq \delta^k + 3\sum_{i=1}^p \bar{\rho}_i^2 P_i^k + 3\sum_{i=1}^p \bar{\rho}_i^2 \tilde{S}_i^k - \frac{1}{4}\sum_{i=1}^p \gamma_i \mathbb{E}\left[\left\| \nabla_i f(X^k) \right\|_{(i)\star}^2 \right]$$

$$
\begin{aligned}
& -\sum_{i=1}^{p}\left(\frac{1}{4\gamma_i}-\frac{L_i^0}{2}\right)\gamma_i^2\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right], && (17)\\
R_i^{k+1} \;\; \leq \;\; & \left(1-\frac{\alpha_P}{2}\right)R_i^k+\frac{2}{\alpha_P}\gamma_i^3\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right], && (18)\\
\tilde{S}_i^{k+1} \;\; \leq \;\; & \left(1-\frac{\alpha_D}{2}\right)\tilde{S}_i^k+\frac{6\beta_i^2}{\alpha_D}\tilde{P}_i^k+\frac{6\beta_i^2(\tilde{L}_i^0)^2}{\alpha_D\underline{\rho}_i^2}\gamma_i^3\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right] \\
& +\frac{6\beta_i^2(\tilde{L}_i^0)^2}{\alpha_D\underline{\rho}_i^2}R_i^{k+1}+(1-\alpha_D)\sigma_i^2\beta_i^2\gamma_i, && (19)\\
\tilde{P}_i^{k+1} \;\; \leq \;\; & \left(1-\frac{\beta_i}{2}\right)\tilde{P}_i^k+\frac{2(\tilde{L}_i^0)^2}{\beta_i\underline{\rho}_i^2}\gamma_i^3\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right] \\
& +\frac{\beta_i^2}{\underline{\rho}_i^2}\left(1+\frac{2}{\beta_i}\right)(\tilde{L}_i^0)^2R_i^{k+1}+\sigma_i^2\beta_i^2\gamma_i, && (20)\\
P_i^{k+1} \;\; \leq \;\; & \left(1-\frac{\beta_i}{2}\right)P_i^k+\frac{2(L_i^0)^2}{\beta_i\underline{\rho}_i^2}\gamma_i^3\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right] \\
& +\frac{\beta_i^2}{\underline{\rho}_i^2}\left(1+\frac{2}{\beta_i}\right)(L_i^0)^2R_i^{k+1}+\frac{\sigma_i^2\beta_i^2\gamma_i}{n}. && (21)
\end{aligned}
$$

Now, let $A_i, B_i, C_i, D_i > 0$ be some constants to be determined later, and define

$$
\Psi^k \;\; := \;\; \delta^k+\sum_{i=1}^{p}A_iP_i^k+\sum_{i=1}^{p}B_i\tilde{P}_i^k+\sum_{i=1}^{p}C_i\tilde{S}_i^k+\sum_{i=1}^{p}D_iR_i^k.
$$

Then, applying (17), (19), (20), and (21), we have

$$
\begin{aligned}
& \Psi^{k+1} \\
= \;\; & \delta^{k+1}+\sum_{i=1}^{p}A_iP_i^{k+1}+\sum_{i=1}^{p}B_i\tilde{P}_i^{k+1}+\sum_{i=1}^{p}C_i\tilde{S}_i^{k+1}+\sum_{i=1}^{p}D_iR_i^{k+1} \\
\leq \;\; & \delta^k+3\sum_{i=1}^{p}\bar{\rho}_i^2P_i^k+3\sum_{i=1}^{p}\bar{\rho}_i^2\tilde{S}_i^k-\frac{1}{4}\sum_{i=1}^{p}\gamma_i\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}^2\right] \\
& -\sum_{i=1}^{p}\left(\frac{1}{4\gamma_i}-\frac{L_i^0}{2}\right)\gamma_i^2\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right] \\
& +\sum_{i=1}^{p}A_i\left(\left(1-\frac{\beta_i}{2}\right)P_i^k+\frac{2(L_i^0)^2}{\beta_i\underline{\rho}_i^2}\gamma_i^3\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right]+\frac{\beta_i^2}{\underline{\rho}_i^2}\left(1+\frac{2}{\beta_i}\right)(L_i^0)^2R_i^{k+1}+\frac{\sigma_i^2\beta_i^2\gamma_i}{n}\right) \\
& +\sum_{i=1}^{p}B_i\left(\left(1-\frac{\beta_i}{2}\right)\tilde{P}_i^k+\frac{2(\tilde{L}_i^0)^2}{\beta_i\underline{\rho}_i^2}\gamma_i^3\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right]+\frac{\beta_i^2}{\underline{\rho}_i^2}\left(1+\frac{2}{\beta_i}\right)(\tilde{L}_i^0)^2R_i^{k+1}+\sigma_i^2\beta_i^2\gamma_i\right) \\
& +\sum_{i=1}^{p}C_i\left(\left(1-\frac{\alpha_D}{2}\right)\tilde{S}_i^k+\frac{6\beta_i^2}{\alpha_D}\tilde{P}_i^k+\frac{6\beta_i^2(\tilde{L}_i^0)^2}{\alpha_D\underline{\rho}_i^2}\gamma_i^3\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right]\right) \\
& +\sum_{i=1}^{p}C_i\left(\frac{6\beta_i^2(\tilde{L}_i^0)^2}{\alpha_D\underline{\rho}_i^2}R_i^{k+1}+(1-\alpha_D)\sigma_i^2\beta_i^2\gamma_i\right)+\sum_{i=1}^{p}D_iR_i^{k+1}
\end{aligned}
$$

$$
\begin{aligned}
&= \delta^k + \sum_{i=1}^p \left(3\bar{\rho}_i^2 + A_i\left(1-\frac{\beta_i}{2}\right)\right) P_i^k + \sum_{i=1}^p \left(B_i\left(1-\frac{\beta_i}{2}\right) + C_i\frac{6\beta_i^2}{\alpha_D}\right)\tilde{P}_i^k \\
&\quad + \sum_{i=1}^p \left(3\bar{\rho}_i^2 + C_i\left(1-\frac{\alpha_D}{2}\right)\right)\tilde{S}_i^k - \frac{1}{4}\sum_{i=1}^p \gamma_i \mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}^2\right] \\
&\quad + \sum_{i=1}^p \left(A_i\frac{\beta_i^2}{\underline{\rho}_i^2}\left(1+\frac{2}{\beta_i}\right)(L_i^0)^2 + B_i\frac{\beta_i^2}{\underline{\rho}_i^2}\left(1+\frac{2}{\beta_i}\right)(\tilde{L}_i^0)^2 + C_i\frac{6\beta_i^2(\tilde{L}_i^0)^2}{\alpha_D\underline{\rho}_i^2} + D_i\right) R_i^{k+1} \\
&\quad + \sum_{i=1}^p \left(A_i\frac{2(L_i^0)^2}{\beta_i\underline{\rho}_i^2} + B_i\frac{2(\tilde{L}_i^0)^2}{\beta_i\underline{\rho}_i^2} + C_i\frac{6\beta_i^2(\tilde{L}_i^0)^2}{\alpha_D\underline{\rho}_i^2}\right)\gamma_i^3\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right] \\
&\quad - \sum_{i=1}^p \left(\frac{1}{4\gamma_i} - \frac{L_i^0}{2}\right)\gamma_i^2\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right] + \sum_{i=1}^p \left(\frac{A_i}{n} + B_i + C_i(1-\alpha_D)\right)\sigma_i^2\beta_i^2\gamma_i.
\end{aligned}
$$

Then, using (18) gives

$$
\begin{aligned}
&\Psi^{k+1} \\
&\leq \delta^k + \sum_{i=1}^p \left(3\bar{\rho}_i^2 + A_i\left(1-\frac{\beta_i}{2}\right)\right) P_i^k + \sum_{i=1}^p \left(B_i\left(1-\frac{\beta_i}{2}\right) + C_i\frac{6\beta_i^2}{\alpha_D}\right)\tilde{P}_i^k \\
&\quad + \sum_{i=1}^p \left(3\bar{\rho}_i^2 + C_i\left(1-\frac{\alpha_D}{2}\right)\right)\tilde{S}_i^k - \frac{1}{4}\sum_{i=1}^p \gamma_i \mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}^2\right] \\
&\quad + \sum_{i=1}^p \left(A_i\frac{\beta_i^2}{\underline{\rho}_i^2}\left(1+\frac{2}{\beta_i}\right)(L_i^0)^2 + B_i\frac{\beta_i^2}{\underline{\rho}_i^2}\left(1+\frac{2}{\beta_i}\right)(\tilde{L}_i^0)^2 + C_i\frac{6\beta_i^2(\tilde{L}_i^0)^2}{\alpha_D\underline{\rho}_i^2} + D_i\right)\left(1-\frac{\alpha_P}{2}\right) R_i^k \\
&\quad + \sum_{i=1}^p \left(A_i\frac{\beta_i^2}{\underline{\rho}_i^2}\left(1+\frac{2}{\beta_i}\right)(L_i^0)^2 + B_i\frac{\beta_i^2}{\underline{\rho}_i^2}\left(1+\frac{2}{\beta_i}\right)(\tilde{L}_i^0)^2 + C_i\frac{6\beta_i^2(\tilde{L}_i^0)^2}{\alpha_D\underline{\rho}_i^2} + D_i\right)\frac{2}{\alpha_P}\gamma_i^3\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right] \\
&\quad + \sum_{i=1}^p \left(A_i\frac{2(L_i^0)^2}{\beta_i\underline{\rho}_i^2} + B_i\frac{2(\tilde{L}_i^0)^2}{\beta_i\underline{\rho}_i^2} + C_i\frac{6\beta_i^2(\tilde{L}_i^0)^2}{\alpha_D\underline{\rho}_i^2}\right)\gamma_i^3\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right] \\
&\quad - \sum_{i=1}^p \left(\frac{1}{4\gamma_i} - \frac{L_i^0}{2}\right)\gamma_i^2\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right] + \sum_{i=1}^p \left(\frac{A_i}{n} + B_i + C_i(1-\alpha_D)\right)\sigma_i^2\beta_i^2\gamma_i.
\end{aligned}
$$

Taking $A_i = \frac{6\bar{\rho}_i^2}{\beta_i}$, $B_i = \frac{72\bar{\rho}_i^2\beta_i}{\alpha_D^2}$, $C_i = \frac{6\bar{\rho}_i^2}{\alpha_D}$ and

$$
\begin{aligned}
D_i &= \left(A_i\frac{\beta_i^2}{\underline{\rho}_i^2}\left(1+\frac{2}{\beta_i}\right)(L_i^0)^2 + B_i\frac{\beta_i^2}{\underline{\rho}_i^2}\left(1+\frac{2}{\beta_i}\right)(\tilde{L}_i^0)^2 + C_i\frac{6\beta_i^2(\tilde{L}_i^0)^2}{\alpha_D\underline{\rho}_i^2}\right)\left(\frac{2}{\alpha_P} - 1\right) \\
&= \frac{6\bar{\rho}_i^2}{\underline{\rho}_i^2}\left((\beta_i+2)(L_i^0)^2 + \frac{6\beta_i^2(2\beta_i+5)}{\alpha_D^2}(\tilde{L}_i^0)^2\right)\left(\frac{2}{\alpha_P} - 1\right),
\end{aligned}
$$

we obtain

$$
3\bar{\rho}_i^2 + A_i\left(1-\frac{\beta_i}{2}\right) = 3\bar{\rho}_i^2 + \frac{6\bar{\rho}_i^2}{\beta_i}\left(1-\frac{\beta_i}{2}\right) = A_i,
$$

$$
\begin{aligned}
B_i\left(1-\frac{\beta_i}{2}\right)+C_i\frac{6\beta_i^2}{\alpha_D} &= \frac{72\bar{\rho}_i^2\beta_i}{\alpha_D^2}\left(1-\frac{\beta_i}{2}\right)+\frac{6\bar{\rho}_i^2}{\alpha_D}\frac{6\beta_i^2}{\alpha_D}=B_i,\\
3\bar{\rho}_i^2+C_i\left(1-\frac{\alpha_D}{2}\right) &= 3\bar{\rho}_i^2+\frac{6\bar{\rho}_i^2}{\alpha_D}\left(1-\frac{\alpha_D}{2}\right)=C_i,
\end{aligned}
$$

and

$$
\begin{aligned}
&\left(A_i\frac{\beta_i^2}{\underline{\rho}_i^2}\left(1+\frac{2}{\beta_i}\right)(L_i^0)^2+B_i\frac{\beta_i^2}{\underline{\rho}_i^2}\left(1+\frac{2}{\beta_i}\right)(\tilde{L}_i^0)^2+C_i\frac{6\beta_i^2(\tilde{L}_i^0)^2}{\alpha_D\underline{\rho}_i^2}+D_i\right)\left(1-\frac{\alpha_P}{2}\right)\\
&= \quad \left(\frac{D_i}{\frac{2}{\alpha_P}-1}+D_i\right)\left(1-\frac{\alpha_P}{2}\right)=D_i.
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
&\Psi^{k+1}\\
&\leq \quad \delta^k+\sum_{i=1}^p A_iP_i^k+\sum_{i=1}^p B_i\tilde{P}_i^k+\sum_{i=1}^p C_i\tilde{S}_i^k+\sum_{i=1}^p D_iR_i^k-\frac{1}{4}\sum_{i=1}^p\gamma_i\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}^2\right]\\
&\quad +\sum_{i=1}^p\left(\frac{D_i}{\frac{2}{\alpha_P}-1}+D_i\right)\frac{2}{\alpha_P}\gamma_i^3\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right]-\sum_{i=1}^p\left(\frac{1}{4\gamma_i}-\frac{L_i^0}{2}\right)\gamma_i^2\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right]\\
&\quad +\sum_{i=1}^p\left(\frac{6\bar{\rho}_i^2}{\beta_i}\frac{2(L_i^0)^2}{\beta_i\underline{\rho}_i^2}+\frac{72\bar{\rho}_i^2\beta_i}{\alpha_D^2}\frac{2(\tilde{L}_i^0)^2}{\beta_i\underline{\rho}_i^2}+\frac{6\bar{\rho}_i^2}{\alpha_D}\frac{6\beta_i^2(\tilde{L}_i^0)^2}{\alpha_D\underline{\rho}_i^2}\right)\gamma_i^3\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right]\\
&\quad +\sum_{i=1}^p\left(\frac{1}{n}\frac{6\bar{\rho}_i^2}{\beta_i}+\frac{72\bar{\rho}_i^2\beta_i}{\alpha_D^2}+\frac{6\bar{\rho}_i^2}{\alpha_D}(1-\alpha_D)\right)\sigma_i^2\beta_i^2\gamma_i\\
&= \quad \Psi^k-\frac{1}{4}\sum_{i=1}^p\gamma_i\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}^2\right]-\sum_{i=1}^p\left(\frac{1}{4\gamma_i}-\frac{L_i^0}{2}\right)\gamma_i^2\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right]\\
&\quad +\sum_{i=1}^p\left(\frac{12\bar{\rho}_i^2}{\beta_i^2\underline{\rho}_i^2}(L_i^0)^2+\frac{144\bar{\rho}_i^2}{\alpha_D^2\underline{\rho}_i^2}(\tilde{L}_i^0)^2+\frac{36\beta_i^2\bar{\rho}_i^2}{\alpha_D^2\underline{\rho}_i^2}(\tilde{L}_i^0)^2+\frac{4D_i}{\alpha_P(2-\alpha_P)}\right)\gamma_i^3\mathbb{E}\left[\left\|G_i^k\right\|_{(i)\star}^2\right]\\
&\quad +6\sum_{i=1}^p\left(\frac{1}{n}+\frac{12\beta_i^2}{\alpha_D^2}+\frac{(1-\alpha_D)\beta_i}{\alpha_D}\right)\sigma_i^2\bar{\rho}_i^2\beta_i\gamma_i.
\end{aligned}
$$

Now, note that

$$
\begin{aligned}
&\frac{1}{4\gamma_i}-\frac{L_i^0}{2}-\gamma_i\left(\frac{12\bar{\rho}_i^2}{\beta_i^2\underline{\rho}_i^2}(L_i^0)^2+\frac{144\bar{\rho}_i^2}{\alpha_D^2\underline{\rho}_i^2}(\tilde{L}_i^0)^2+\frac{36\beta_i^2\bar{\rho}_i^2}{\alpha_D^2\underline{\rho}_i^2}(\tilde{L}_i^0)^2+\frac{4D_i}{\alpha_P(2-\alpha_P)}\right)\\
&=\frac{1}{4\gamma_i}-\frac{L_i^0}{2}\\
&\quad -\gamma_i\underbrace{\frac{\bar{\rho}_i^2}{\underline{\rho}_i^2}\left(\frac{12}{\beta_i^2}(L_i^0)^2+\frac{24\left(\beta_i+2\right)}{\alpha_P^2}(L_i^0)^2+\frac{36\left(\beta_i^2+4\right)}{\alpha_D^2}(\tilde{L}_i^0)^2+\frac{144\beta_i^2\left(2\beta_i+5\right)}{\alpha_P^2\alpha_D^2}(\tilde{L}_i^0)^2\right)}_{:=\zeta_i}\geq 0
\end{aligned}
$$

for $\gamma_i \leq \frac{1}{2\sqrt{\zeta_i}+2L_i^0}$. For such a choice of the stepsizes, we have

$$\Psi^{k+1} \leq \Psi^k - \frac{1}{4}\sum_{i=1}^p \gamma_i \mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}^2\right] + \sum_{i=1}^p \underbrace{6\left(\frac{1}{n} + \frac{12\beta_i^2}{\alpha_D^2} + \frac{(1-\alpha_D)\beta_i}{\alpha_D}\right)\sigma_i^2\bar{\rho}_i^2\beta_i}_{:=\xi_i}\gamma_i.$$

Summing over the first K iterations gives

$$\sum_{k=0}^{K-1}\sum_{i=1}^p \gamma_i \mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}^2\right] \leq 4\sum_{k=0}^{K-1}\left(\Psi^k - \Psi^{k+1}\right) + 4\sum_{k=0}^{K-1}\sum_{i=1}^p \xi_i\gamma_i \leq 4\Psi^0 + 4K\sum_{i=1}^p \xi_i\gamma_i,$$

and lastly, dividing by $\frac{K}{p}\sum_{l=1}^p \gamma_l$, we obtain

$$\frac{1}{K}\sum_{k=0}^{K-1}\sum_{i=1}^p \frac{\gamma_i}{\frac{1}{p}\sum_{l=1}^p \gamma_l}\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}^2\right] \leq \frac{4\Psi^0 p}{K\sum_{l=1}^p \gamma_l} + \frac{4\sum_{i=1}^p \xi_i\gamma_i}{\frac{1}{p}\sum_{i=1}^p \gamma_i}.$$

Substituting $X_i^0 = W_i^0$ proves the theorem statement. ■

Proof [Proof of Theorem 31] Substituting the choice of γ from (15) in (14), we have

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(X^k)\right\|_\star^2\right] &\leq \frac{4\Psi^0}{K\gamma_1} + 24\left(\frac{1}{n} + \frac{(1-\alpha_D)\beta_1}{\alpha_D} + \frac{12\beta_1^2}{\alpha_D^2}\right)\sigma_1^2\bar{\rho}_1^2\beta_1 \\ &= \mathcal{O}\left(\frac{\Psi^0\bar{\rho}_1^2\tilde{L}_1^0}{\underline{\rho}_1^2\alpha_P\alpha_D K} + \frac{\Psi^0\bar{\rho}_1^2 L_1^0}{\underline{\rho}_1^2\beta_1 K} + \frac{\bar{\rho}_1^2\beta_1\sigma_1^2}{n} + \frac{\bar{\rho}_1^2\beta_1^2\sigma_1^2}{\alpha_D} + \frac{\bar{\rho}_1^2\beta_1^3\sigma_1^2}{\alpha_D^2}\right).\end{aligned}$$

Then, choosing β_1 as in (16) guarantees that $\frac{\bar{\rho}_1^2\beta_1\sigma_1^2}{n}, \frac{\bar{\rho}_1^2\beta_1^2\sigma_1^2}{\alpha_D}, \frac{\bar{\rho}_1^2\beta_1^3\sigma_1^2}{\alpha_D^2} \leq \frac{\Psi^0\bar{\rho}_1^2 L_1^0}{\underline{\rho}_1^2\beta_1 K}$. Substituting this into the upper bound gives

$$\begin{aligned}&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(X^k)\right\|_\star^2\right] \\ &\leq \mathcal{O}\left(\frac{\Psi^0\bar{\rho}_1^2\tilde{L}_1^0}{\underline{\rho}_1^2\alpha_P\alpha_D K} + \left(\frac{\Psi^0\bar{\rho}_1^4\sigma_1^2 L_1^0}{\underline{\rho}_1^2 n K}\right)^{1/2} + \left(\frac{\Psi^0\bar{\rho}_1^3\sigma_1 L_1^0}{\underline{\rho}_1^2\sqrt{\alpha_D}K}\right)^{2/3} + \left(\frac{\Psi^0\underline{\rho}_1^{8/3}\sigma_1^{2/3}L_1^0}{\bar{\rho}_1^2\alpha_D^{2/3}K}\right)^{3/4}\right)\end{aligned}$$

as needed. ■

F.4.2. LAYER-WISE (L^0, L^1)–SMOOTH REGIME

As in Appendix F.3.2, in the generalized smooth setting we consider EF21-Muon without primal compression.

Theorem 36 *Let Assumptions 1, 2, 8, 9 and 10 hold. Let $\{X^k\}_{k=0}^{K-1}$, $K \geq 1$, be the iterates of Algorithm 3 run with $\mathcal{C}_i^k \equiv \mathcal{I}$ (the identity compressor), $\mathcal{C}_{i,j}^k \in \mathbb{B}_2(\alpha_D)$, $\beta_i \equiv \beta = \frac{1}{(K+1)^{1/2}}$ and*

$$0 \leq t_i^k \equiv t_i = \frac{\eta_i}{(K+1)^{3/4}}, \qquad i = 1, \ldots, p,$$

where $\eta_i^2 \leq \min\left\{\frac{(K+1)^{1/2}}{6(L_i^1)^2}, \frac{(1-\sqrt{1-\alpha_D})\underline{\rho}_i(K+1)^{1/2}}{24\sqrt{1-\alpha_D}\bar{\rho}_i(L_{i,\max}^1)^2}, \frac{\beta_{\min}\underline{\rho}_i(K+1)^{1/2}}{24\bar{\rho}_i(L_{i,\max}^1)^2}, 1\right\}$. Then

$$\begin{aligned}
&\min_{k=0,\ldots,K} \sum_{i=1}^p \frac{\eta_i}{\frac{1}{p}\sum_{l=1}^p \eta_l} \mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right] \\
\leq \quad & \frac{3\Psi^0}{(K+1)^{1/4}\frac{1}{p}\sum_{l=1}^p \eta_l} + \frac{6}{(K+1)^{1/2}} \sum_{i=1}^p \frac{\eta_i \bar{\rho}_i}{\frac{1}{p}\sum_{l=1}^p \eta_l} \mathbb{E}\left[\left\|\nabla_i f(X^0) - M_i^0\right\|_2\right] \\
& + \left(\frac{8}{(K+1)^{1/4}} + \frac{8\sqrt{1-\alpha_D}}{(1-\sqrt{1-\alpha_D})(K+1)^{3/4}}\right) \frac{1}{n} \sum_{j=1}^n \frac{\max_{i\in[p]} \eta_i^2 \frac{\bar{\rho}_i}{\underline{\rho}_i} (L_{i,j}^1)^2}{\frac{1}{p}\sum_{l=1}^p \eta_l} \left(f^\star - f_j^\star\right) \\
& + \sum_{i=1}^p \frac{\eta_i^2}{\frac{1}{p}\sum_{l=1}^p \eta_l} \left(\frac{L_i^0}{(K+1)^{3/4}} + \frac{4\bar{\rho}_i \bar{L}_i^0}{\underline{\rho}_i (K+1)^{1/4}} + \frac{4\bar{\rho}_i\sqrt{1-\alpha_D}\bar{L}_i^0}{\underline{\rho}_i(1-\sqrt{1-\alpha_D})(K+1)^{3/4}}\right) \\
& + \sum_{i=1}^p \frac{\eta_i \bar{\rho}_i \sigma_i}{\frac{1}{p}\sum_{l=1}^p \eta_l} \left(\frac{4\sqrt{1-\alpha_D}}{(1-\sqrt{1-\alpha_D})(K+1)^{1/2}} + \frac{2}{\sqrt{n}(K+1)^{1/4}}\right),
\end{aligned}$$

where $M_i^0 := \frac{1}{n}\sum_{j=1}^n M_{i,j}^0$ and

$$\begin{aligned}
\Psi^0 \quad := \quad & f(X^0) - f^\star + \sum_{i=1}^p \frac{2t_i\bar{\rho}_i}{1-\sqrt{1-\alpha_D}} \frac{1}{n} \sum_{j=1}^n \mathbb{E}\left[\left\|M_{i,j}^0 - G_{i,j}^0\right\|_2\right] \\
& + \sum_{i=1}^p \frac{2t_i\bar{\rho}_i\sqrt{1-\alpha_D}}{1-\sqrt{1-\alpha_D}} \frac{1}{n} \sum_{j=1}^n \mathbb{E}\left[\left\|\nabla_i f_j(X^0) - M_{i,j}^0\right\|_2\right].
\end{aligned}$$

Corollary 37 *Let the assumptions of Theorem 36 hold and let $\{X^k\}_{k=0}^{K-1}$, $K \geq 1$, be the iterates of Algorithm 3 initialized with $M_{i,j}^0 = \nabla_i f_j(X^0; \xi_j^0)$, $G_{i,j}^0 = \mathcal{C}_{i,j}^0(\nabla_i f_j(X^0; \xi_j^0))$, $j \in [n]$, and run with $\mathcal{C}_i^k \equiv \mathcal{I}$ (the identity compressor), $\mathcal{C}_{i,j}^k \in \mathbb{B}_2(\alpha_D)$, $\beta_i \equiv \beta = \frac{1}{(K+1)^{1/2}}$ and*

$$0 \leq t_i^k \equiv t_i = \frac{\eta_i}{(K+1)^{3/4}}, \qquad i = 1, \ldots, p,$$

where $\eta_i^2 \leq \min\left\{\frac{(K+1)^{1/2}}{6(L_i^1)^2}, \frac{(1-\sqrt{1-\alpha_D})\underline{\rho}_i(K+1)^{1/2}}{24\sqrt{1-\alpha_D}\bar{\rho}_i(L_{i,\max}^1)^2}, \frac{\beta_{\min}\underline{\rho}_i(K+1)^{1/2}}{24\bar{\rho}_i(L_{i,\max}^1)^2}, 1\right\}$. Then, the result in Theorem 30 guarantees that

$$\begin{aligned}
&\min_{k=0,\ldots,K} \sum_{i=1}^p \frac{\eta_i}{\frac{1}{p}\sum_{l=1}^p \eta_l} \mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right] \\
\leq \quad & \frac{3}{(K+1)^{1/4}\frac{1}{p}\sum_{l=1}^p \eta_l} \left(f(X^0) - f^\star + \sum_{i=1}^p \frac{4\sqrt{1-\alpha_D}\eta_i\bar{\rho}_i\sigma_i}{(K+1)^{3/4}(1-\sqrt{1-\alpha_D})}\right.
\end{aligned}$$

$$
\begin{aligned}
&+\frac{6}{(K+1)^{1/2}}\sum_{i=1}^{p}\frac{\bar{\rho}_i\eta_i\sigma_i}{\sqrt{n}\frac{1}{p}\sum_{l=1}^{p}\eta_l}\\
&+\left(\frac{8}{(K+1)^{1/4}}+\frac{8\sqrt{1-\alpha_D}}{(K+1)^{3/4}(1-\sqrt{1-\alpha_D})}\right)\frac{1}{n}\sum_{j=1}^{n}\frac{\max_{i\in[p]}\eta_i^2\frac{\bar{\rho}_i}{\underline{\rho}_i}(L^1_{i,j})^2}{\frac{1}{p}\sum_{l=1}^{p}\eta_l}\left(f^\star-f_j^\star\right)\\
&+\sum_{i=1}^{p}\frac{\eta_i^2}{\frac{1}{p}\sum_{l=1}^{p}\eta_l}\left(\frac{L_i^0}{(K+1)^{3/4}}+\frac{4\bar{\rho}_i\bar{L}_i^0}{\underline{\rho}_i(K+1)^{1/4}}+\frac{4\bar{\rho}_i\sqrt{1-\alpha_D}\bar{L}_i^0}{\underline{\rho}_i(K+1)^{3/4}(1-\sqrt{1-\alpha_D})}\right)\\
&+\sum_{i=1}^{p}\frac{\eta_i\bar{\rho}_i\sigma_i}{\frac{1}{p}\sum_{l=1}^{p}\eta_l}\left(\frac{4\sqrt{1-\alpha_D}}{(K+1)^{1/2}(1-\sqrt{1-\alpha_D})}+\frac{2}{\sqrt{n}(K+1)^{1/4}}\right).
\end{aligned}
$$

Remark 38 *Theorem 6 follows from Theorem 37 by setting $p=1$:*

$$
\begin{aligned}
&\min_{k=0,\ldots,K}\mathbb{E}\left[\left\|\nabla f(X^k)\right\|_\star\right]\\
\leq\;&\frac{3\left(f(X^0)-f^\star\right)}{\eta(K+1)^{1/4}}+\frac{12\sqrt{1-\alpha_D}\bar{\rho}\sigma}{(1-\sqrt{1-\alpha_D})(K+1)}+\frac{6\bar{\rho}\sigma}{\sqrt{n}(K+1)^{1/2}}\\
&+\frac{\eta\bar{\rho}}{\underline{\rho}}\left(\frac{8}{(K+1)^{1/4}}+\frac{8\sqrt{1-\alpha_D}}{(1-\sqrt{1-\alpha_D})(K+1)^{3/4}}\right)\frac{1}{n}\sum_{j=1}^{n}(L_j^1)^2\left(f^\star-f_j^\star\right)\\
&+\frac{\eta L^0}{(K+1)^{3/4}}+\frac{\eta\bar{\rho}}{\underline{\rho}}\left(\frac{4}{(K+1)^{1/4}}+\frac{4\sqrt{1-\alpha_D}}{(1-\sqrt{1-\alpha_D})(K+1)^{3/4}}\right)\bar{L}^0\\
&+\frac{4\bar{\rho}\sigma\sqrt{1-\alpha_D}}{(1-\sqrt{1-\alpha_D})(K+1)^{1/2}}+\frac{2\bar{\rho}\sigma}{\sqrt{n}(K+1)^{1/4}}\\
\leq\;&\frac{3\left(f(X^0)-f^\star\right)}{\eta(K+1)^{1/4}}+\frac{16\sqrt{1-\alpha_D}\bar{\rho}\sigma}{(1-\sqrt{1-\alpha_D})(K+1)^{1/2}}+\frac{\eta L^0}{(K+1)^{3/4}}+\frac{8\bar{\rho}\sigma}{\sqrt{n}(K+1)^{1/4}}\\
&+\frac{\eta\bar{\rho}}{\underline{\rho}}\left(\frac{8}{(K+1)^{1/4}}+\frac{8\sqrt{1-\alpha_D}}{(1-\sqrt{1-\alpha_D})(K+1)^{3/4}}\right)\left(\frac{1}{n}\sum_{j=1}^{n}(L_j^1)^2\left(f^\star-f_j^\star\right)+\bar{L}^0\right).
\end{aligned}
$$

Proof [Proof of Theorem 36] By Theorem 15 and Jensen's inequality

$$
\begin{aligned}
f(X^{k+1})\;\leq\;&f(X^k)+\sum_{i=1}^{p}2t_i\left\|\nabla_i f(X^k)-G_i^k\right\|_{(i)\star}-\sum_{i=1}^{p}t_i\left\|\nabla_i f(X^k)\right\|_{(i)\star}\\
&+\sum_{i=1}^{p}\frac{L_i^0+L_i^1\left\|\nabla_i f(X^k)\right\|_{(i)\star}}{2}t_i^2\\
\leq\;&f(X^k)+\sum_{i=1}^{p}\left(2t_i\left\|\nabla_i f(X^k)-M_i^k\right\|_{(i)\star}+2t_i\left\|M_i^k-G_i^k\right\|_{(i)\star}\right)\\
&-\sum_{i=1}^{p}t_i\left\|\nabla_i f(X^k)\right\|_{(i)\star}+\sum_{i=1}^{p}\left(\frac{L_i^0}{2}t_i^2+\frac{L_i^1}{2}\left\|\nabla_i f(X^k)\right\|_{(i)\star}t_i^2\right)\\
\leq\;&f(X^k)+\sum_{i=1}^{p}\left(2\bar{\rho}_i t_i\left\|\nabla_i f(X^k)-M_i^k\right\|_2+2\bar{\rho}_i t_i\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|M_{i,j}^k-G_{i,j}^k\right\|_2\right]\right)
\end{aligned}
$$

$$-\sum_{i=1}^{p} t_i \left\|\nabla_i f(X^k)\right\|_{(i)\star} + \sum_{i=1}^{p} \left(\frac{L_i^0}{2} t_i^2 + \frac{L_i^1}{2} \left\|\nabla_i f(X^k)\right\|_{(i)\star} t_i^2\right).$$

To simplify the notation, let $\delta^k := \mathbb{E}\left[f(X^k) - f^\star\right]$, $P_i^k := \mathbb{E}\left[\left\|\nabla_i f(X^k) - M_i^k\right\|_2\right]$, $\tilde{P}_i^k := \frac{1}{n}\sum_{j=1}^n \mathbb{E}\left[\left\|\nabla_i f_j(X^k) - M_{i,j}^k\right\|_2\right]$ and $\tilde{S}_i^k := \frac{1}{n}\sum_{j=1}^n \mathbb{E}\left[\left\|M_{i,j}^k - G_{i,j}^k\right\|_2\right]$. Then, Lemmas 20, 21, and the descent inequality above yield

$$\begin{aligned}
\tilde{S}_i^{k+1} &\overset{(20)}{\leq} \sqrt{1-\alpha_D}\tilde{S}_i^k + \sqrt{1-\alpha_D}\beta_i \tilde{P}_i^k + \frac{t_i\sqrt{1-\alpha_D}\beta_i}{\underline{\rho}_i}\left(\frac{1}{n}\sum_{j=1}^n L_{i,j}^1 \mathbb{E}\left[\left\|\nabla_i f_j(X^k)\right\|_{(i)\star}\right]\right) \\
&\quad + \frac{t_i\sqrt{1-\alpha_D}\beta_i \bar{L}_i^0}{\underline{\rho}_i} + \sqrt{1-\alpha_D}\beta_i\sigma_i, && (22) \\
P_i^k &\overset{(21)}{\leq} (1-\beta_i)^k P_i^0 + \frac{t_i}{\underline{\rho}_i}\frac{1}{n}\sum_{j=1}^n L_{i,j}^1 \sum_{l=0}^{k-1}(1-\beta_i)^{k-l}\mathbb{E}\left[\left\|\nabla_i f_j(X^l)\right\|_{(i)\star}\right] \\
&\quad + \frac{t_i \bar{L}_i^0}{\underline{\rho}_i \beta_i} + \sigma_i\sqrt{\frac{\beta_i}{n}}, && (23) \\
\tilde{P}_i^{k+1} &\overset{(21)}{\leq} (1-\beta_i)\tilde{P}_i^k + \frac{t_i(1-\beta_i)}{\underline{\rho}_i}\frac{1}{n}\sum_{j=1}^n L_{i,j}^1 \mathbb{E}\left[\left\|\nabla_i f_j(X^k)\right\|_{(i)\star}\right] \\
&\quad + \frac{t_i(1-\beta_i)\bar{L}_i^0}{\underline{\rho}_i} + \beta_i\sigma_i, && (24) \\
\delta^{k+1} &\leq \delta^k - \sum_{i=1}^p t_i \mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right] && (25) \\
&\quad + \sum_{i=1}^p \left(2t_i\bar{\rho}_i P_i^k + 2t_i\bar{\rho}_i\tilde{S}_i^k + \frac{t_i^2 L_i^0}{2} + \frac{t_i^2 L_i^1}{2}\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right]\right).
\end{aligned}$$

Let $A_i, B_i > 0$ be some constants to be determined later, and define

$$\Psi^k := \delta^k + \sum_{i=1}^p A_i \tilde{S}_i^k + \sum_{i=1}^p B_i \tilde{P}_i^k.$$

Then, using (22), (24) and (25)

$$\begin{aligned}
&\Psi^{k+1} \\
&= \delta^{k+1} + \sum_{i=1}^p A_i \tilde{S}_i^{k+1} + \sum_{i=1}^p B_i \tilde{P}_i^{k+1} \\
&\leq \delta^k - \sum_{i=1}^p t_i \mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right] + \sum_{i=1}^p \left(2t_i\bar{\rho}_i P_i^k + 2t_i\bar{\rho}_i\tilde{S}_i^k + \frac{t_i^2 L_i^0}{2} + \frac{t_i^2 L_i^1}{2}\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right]\right) \\
&\quad + \sum_{i=1}^p A_i \left(\sqrt{1-\alpha_D}\tilde{S}_i^k + \sqrt{1-\alpha_D}\beta_i\tilde{P}_i^k + \frac{t_i\sqrt{1-\alpha_D}\beta_i}{\underline{\rho}_i}\left(\frac{1}{n}\sum_{j=1}^n L_{i,j}^1 \mathbb{E}\left[\left\|\nabla_i f_j(X^k)\right\|_{(i)\star}\right]\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{i=1}^{p} A_i\left(\frac{t_i\sqrt{1-\alpha_D}\beta_i\bar{L}_i^0}{\underline{\rho}_i}+\sqrt{1-\alpha_D}\beta_i\sigma_i\right)\\
&+\sum_{i=1}^{p} B_i\left((1-\beta_i)\tilde{P}_i^k+\frac{t_i(1-\beta_i)}{\underline{\rho}_i}\left(\frac{1}{n}\sum_{j=1}^{n}L_{i,j}^1\mathbb{E}\left[\left\|\nabla_i f_j(X^k)\right\|_{(i)\star}\right]\right)+\frac{t_i(1-\beta_i)\bar{L}_i^0}{\underline{\rho}_i}+\beta_i\sigma_i\right)\\
= \quad &\delta^k-\sum_{i=1}^{p}t_i\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right]+\sum_{i=1}^{p}\left(2t_i\bar{\rho}_i+A_i\sqrt{1-\alpha_D}\right)\tilde{S}_i^k\\
&+\sum_{i=1}^{p}\left(A_i\sqrt{1-\alpha_D}\beta_i+B_i(1-\beta_i)\right)\tilde{P}_i^k+\sum_{i=1}^{p}2t_i\bar{\rho}_iP_i^k+\sum_{i=1}^{p}\frac{t_i^2L_i^1}{2}\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right]\\
&+\sum_{i=1}^{p}\frac{t_i}{\underline{\rho}_i}\left(A_i\sqrt{1-\alpha_D}\beta_i+B_i(1-\beta_i)\right)\left(\frac{1}{n}\sum_{j=1}^{n}L_{i,j}^1\mathbb{E}\left[\left\|\nabla_i f_j(X^k)\right\|_{(i)\star}\right]\right)\\
&+\sum_{i=1}^{p}\frac{t_i^2L_i^0}{2}+\sum_{i=1}^{p}A_i\frac{t_i\sqrt{1-\alpha_D}\beta_i\bar{L}_i^0}{\underline{\rho}_i}+\sum_{i=1}^{p}B_i\frac{t_i(1-\beta_i)\bar{L}_i^0}{\underline{\rho}_i}+\sum_{i=1}^{p}A_i\sqrt{1-\alpha_D}\beta_i\sigma_i+\sum_{i=1}^{p}B_i\beta_i\sigma_i.
\end{aligned}
$$

Taking $A_i=\frac{2t_i\bar{\rho}_i}{1-\sqrt{1-\alpha_D}}$ and $B_i=A_i\sqrt{1-\alpha_D}=\frac{2t_i\bar{\rho}\sqrt{1-\alpha_D}}{1-\sqrt{1-\alpha_D}}$, we obtain

$$
\begin{aligned}
2t_i\bar{\rho}+A_i\sqrt{1-\alpha_D}=2t_i\bar{\rho}+\frac{2t_i\bar{\rho}}{1-\sqrt{1-\alpha_D}}\sqrt{1-\alpha_D}=A_i,\\
A_i\sqrt{1-\alpha_D}\beta_i+B_i(1-\beta_i)=A_i\sqrt{1-\alpha_D}\beta_i+A_i\sqrt{1-\alpha_D}(1-\beta_i)=B_i.
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
&\Psi^{k+1}\\
\leq \quad &\delta^k-\sum_{i=1}^{p}t_i\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right]+\sum_{i=1}^{p}A_i\tilde{S}_i^k+\sum_{i=1}^{p}B_i\tilde{P}_i^k+\sum_{i=1}^{p}2t_i\bar{\rho}_iP_i^k\\
&+\sum_{i=1}^{p}\frac{t_i^2L_i^1}{2}\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right]+\sum_{i=1}^{p}\frac{2t_i^2\bar{\rho}_i\sqrt{1-\alpha_D}}{\underline{\rho}_i(1-\sqrt{1-\alpha_D})}\left(\frac{1}{n}\sum_{j=1}^{n}L_{i,j}^1\mathbb{E}\left[\left\|\nabla_i f_j(X^k)\right\|_{(i)\star}\right]\right)\\
&+\sum_{i=1}^{p}\frac{t_i^2L_i^0}{2}+\sum_{i=1}^{p}\frac{2t_i^2\bar{\rho}_i\sqrt{1-\alpha_D}\bar{L}_i^0}{\underline{\rho}_i(1-\sqrt{1-\alpha_D})}+\sum_{i=1}^{p}\frac{4t_i\bar{\rho}_i\sqrt{1-\alpha_D}\beta_i\sigma_i}{1-\sqrt{1-\alpha_D}}\\
\overset{(23)}{\leq} \quad &\delta^k-\sum_{i=1}^{p}t_i\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right]+\sum_{i=1}^{p}A_i\tilde{S}_i^k+\sum_{i=1}^{p}B_i\tilde{P}_i^k\\
&+\sum_{i=1}^{p}2t_i\bar{\rho}_i\left((1-\beta_i)^kP_i^0+\frac{t_i}{\underline{\rho}_i}\frac{1}{n}\sum_{j=1}^{n}L_{i,j}^1\sum_{l=0}^{k-1}(1-\beta_i)^{k-l}\mathbb{E}\left[\left\|\nabla_i f_j(X^l)\right\|_{(i)\star}\right]+\frac{t_i\bar{L}_i^0}{\underline{\rho}_i\beta_i}+\sigma_i\sqrt{\frac{\beta_i}{n}}\right)\\
&+\sum_{i=1}^{p}\frac{t_i^2L_i^1}{2}\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right]+\sum_{i=1}^{p}\frac{2t_i^2\bar{\rho}_i\sqrt{1-\alpha_D}}{\underline{\rho}_i(1-\sqrt{1-\alpha_D})}\left(\frac{1}{n}\sum_{j=1}^{n}L_{i,j}^1\mathbb{E}\left[\left\|\nabla_i f_j(X^k)\right\|_{(i)\star}\right]\right)\\
&+\sum_{i=1}^{p}\frac{t_i^2L_i^0}{2}+\sum_{i=1}^{p}\frac{2t_i^2\bar{\rho}_i\sqrt{1-\alpha_D}\bar{L}_i^0}{\underline{\rho}_i(1-\sqrt{1-\alpha_D})}+\sum_{i=1}^{p}\frac{4t_i\bar{\rho}_i\sqrt{1-\alpha_D}\beta_i\sigma_i}{1-\sqrt{1-\alpha_D}}
\end{aligned}
$$

$$\begin{aligned}
= \quad & \Psi^k - \sum_{i=1}^p t_i \mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right] + \sum_{i=1}^p 2t_i\bar{\rho}_i(1-\beta_i)^k P_i^0 + \frac{1}{2}\sum_{i=1}^p t_i^2 L_i^1 \mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right] \\
& +2\sum_{i=1}^p \frac{t_i^2\bar{\rho}_i}{\underline{\rho}_i}\sum_{l=0}^{k-1}(1-\beta_i)^{k-l}\left(\frac{1}{n}\sum_{j=1}^n L_{i,j}^1 \mathbb{E}\left[\left\|\nabla_i f_j(X^l)\right\|_{(i)\star}\right]\right) \\
& +\frac{2\sqrt{1-\alpha_D}}{1-\sqrt{1-\alpha_D}}\sum_{i=1}^p \frac{t_i^2\bar{\rho}_i}{\underline{\rho}_i}\left(\frac{1}{n}\sum_{j=1}^n L_{i,j}^1 \mathbb{E}\left[\left\|\nabla_i f_j(X^k)\right\|_{(i)\star}\right]\right) + \sum_{i=1}^p \frac{t_i^2 L_i^0}{2} \\
& +\sum_{i=1}^p \frac{2t_i^2\bar{\rho}_i\bar{L}_i^0}{\underline{\rho}_i\beta_i} + \sum_{i=1}^p \frac{2t_i^2\bar{\rho}_i\sqrt{1-\alpha_D}\bar{L}_i^0}{\underline{\rho}_i(1-\sqrt{1-\alpha_D})} + \sum_{i=1}^p \frac{4t_i\bar{\rho}_i\sqrt{1-\alpha_D}\beta_i\sigma_i}{1-\sqrt{1-\alpha_D}} + \sum_{i=1}^p 2t_i\bar{\rho}_i\sigma_i\sqrt{\frac{\beta_i}{n}}. \qquad (26)
\end{aligned}$$

Let us bound the terms involving the norms of the gradients. Using Theorem 23, we get

$$\begin{aligned}
\sum_{i=1}^p t_i^2 L_i^1 \mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right] & \overset{(23)}{\le} 4\max_{i\in[p]}(t_i^2(L_i^1)^2)\mathbb{E}\left[f(X^k)-f^\star\right] + \frac{\sum_{i=1}^p (t_i^2 L_i^1)^2 L_i^0}{\max_{i\in[p]}(t_i^2(L_i^1)^2)} \\
& \le 4\max_{i\in[p]}(t_i^2(L_i^1)^2)\delta^k + \sum_{i=1}^p t_i^2 L_i^0.
\end{aligned}$$

Similarly, Theorem 24 gives

$$\begin{aligned}
& \sum_{i=1}^p \frac{t_i^2\bar{\rho}_i}{\underline{\rho}_i}\sum_{l=0}^{k-1}(1-\beta_i)^{k-l}\frac{1}{n}\sum_{j=1}^n L_{i,j}^1 \mathbb{E}\left[\left\|\nabla_i f_j(X^l)\right\|_{(i)\star}\right] \\
= \quad & \frac{1}{n}\sum_{j=1}^n\sum_{l=0}^{k-1}\sum_{i=1}^p \frac{t_i^2\bar{\rho}_i}{\underline{\rho}_i}(1-\beta_i)^{k-l}L_{i,j}^1 \mathbb{E}\left[\left\|\nabla_i f_j(X^l)\right\|_{(i)\star}\right] \\
\overset{(24)}{\le} \quad & \frac{1}{n}\sum_{j=1}^n\sum_{l=0}^{k-1}\left(4\max_{i\in[p]}\left(\frac{t_i^2\bar{\rho}_i}{\underline{\rho}_i}(1-\beta_i)^{k-l}(L_{i,j}^1)^2\right)\mathbb{E}\left[f_j(X^l)-f^\star\right]\right) \\
& +\frac{1}{n}\sum_{j=1}^n\sum_{l=0}^{k-1}\left(4\max_{i\in[p]}\left(\frac{t_i^2\bar{\rho}_i}{\underline{\rho}_i}(1-\beta_i)^{k-l}(L_{i,j}^1)^2\right)(f^\star - f_j^\star)\right) \\
& +\frac{1}{n}\sum_{j=1}^n\sum_{l=0}^{k-1}\left(\frac{\sum_{i=1}^p\left(\frac{t_i^2\bar{\rho}_i}{\underline{\rho}_i}(1-\beta_i)^{k-l}L_{i,j}^1\right)^2 L_{i,j}^0}{\max_{i\in[p]}\left(\frac{t_i^2\bar{\rho}_i}{\underline{\rho}_i}(1-\beta_i)^{k-l}(L_{i,j}^1)^2\right)}\right) \\
\le \quad & \sum_{l=0}^{k-1} 4\max_{i\in[p],j\in[n]}\left(\frac{t_i^2\bar{\rho}_i}{\underline{\rho}_i}(1-\beta_i)^{k-l}(L_{i,j}^1)^2\right)\delta^l \\
& +\frac{1}{n}\sum_{j=1}^n\sum_{l=0}^{k-1}\left(4\max_{i\in[p]}\left(\frac{t_i^2\bar{\rho}_i}{\underline{\rho}_i}(1-\beta_i)^{k-l}(L_{i,j}^1)^2\right)(f^\star - f_j^\star) + \sum_{i=1}^p\frac{t_i^2\bar{\rho}_i}{\underline{\rho}_i}(1-\beta_i)^{k-l}L_{i,j}^0\right)
\end{aligned}$$

and

$$\sum_{i=1}^p \frac{t_i^2\bar{\rho}_i}{\underline{\rho}_i}\frac{1}{n}\sum_{j=1}^n L_{i,j}^1 \mathbb{E}\left[\left\|\nabla_i f_j(X^k)\right\|_{(i)\star}\right]$$

$$\begin{aligned}
&= \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{p}\frac{t_i^2\bar{\rho}_i L_{i,j}^1}{\underline{\rho}_i}\mathbb{E}\left[\left\|\nabla_i f_j(X^k)\right\|_{(i)\star}\right] \\
&\overset{(24)}{\leq} \frac{1}{n}\sum_{j=1}^{n}\left(4\max_{i\in[p]}\frac{t_i^2\bar{\rho}_i(L_{i,j}^1)^2}{\underline{\rho}_i}\mathbb{E}\left[f_j(X^k)-f^\star\right]+4\max_{i\in[p]}\frac{t_i^2\bar{\rho}_i(L_{i,j}^1)^2}{\underline{\rho}_i}\left(f^\star-f_j^\star\right)\right) \\
&\quad +\frac{1}{n}\sum_{j=1}^{n}\left(\frac{\sum_{i=1}^{p}\left(\frac{t_i^2\bar{\rho}_i}{\underline{\rho}_i}L_{i,j}^1\right)^2L_{i,j}^0}{\max_{i\in[p]}\left(\frac{t_i^2\bar{\rho}_i}{\underline{\rho}_i}(L_{i,j}^1)^2\right)}\right) \\
&\leq 4\max_{i\in[p],j\in[n]}\left(\frac{t_i^2\bar{\rho}_i}{\underline{\rho}_i}(L_{i,j}^1)^2\right)\delta^k+\frac{1}{n}\sum_{j=1}^{n}\left(4\max_{i\in[p]}\frac{t_i^2\bar{\rho}_i(L_{i,j}^1)^2}{\underline{\rho}_i}\left(f^\star-f_j^\star\right)+\sum_{i=1}^{p}\frac{t_i^2\bar{\rho}_iL_{i,j}^0}{\underline{\rho}_i}\right).
\end{aligned}$$

Substituting these bounds in (26), we obtain

$$\begin{aligned}
&\Psi^{k+1} \\
&\leq \Psi^k-\sum_{i=1}^{p}t_i\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right]+\sum_{i=1}^{p}2t_i\bar{\rho}_i(1-\beta_i)^kP_i^0+2\max_{i\in[p]}(t_i^2(L_i^1)^2)\delta^k+\frac{1}{2}\sum_{i=1}^{p}t_i^2L_i^0 \\
&\quad +8\sum_{l=0}^{k-1}\max_{i\in[p],j\in[n]}\left(\frac{t_i^2\bar{\rho}_i}{\underline{\rho}_i}(1-\beta_i)^{k-l}(L_{i,j}^1)^2\right)\delta^l \\
&\quad +2\frac{1}{n}\sum_{j=1}^{n}\sum_{l=0}^{k-1}\left(4\max_{i\in[p]}\left(\frac{t_i^2\bar{\rho}_i}{\underline{\rho}_i}(1-\beta_i)^{k-l}(L_{i,j}^1)^2\right)\left(f^\star-f_j^\star\right)+\sum_{i=1}^{p}\frac{t_i^2\bar{\rho}_i}{\underline{\rho}_i}(1-\beta_i)^{k-l}L_{i,j}^0\right) \\
&\quad +\frac{8\sqrt{1-\alpha_D}}{1-\sqrt{1-\alpha_D}}\max_{i\in[p],j\in[n]}\left(\frac{t_i^2\bar{\rho}_i}{\underline{\rho}_i}(L_{i,j}^1)^2\right)\delta^k \\
&\quad +\frac{2\sqrt{1-\alpha_D}}{1-\sqrt{1-\alpha_D}}\frac{1}{n}\sum_{j=1}^{n}\left(4\max_{i\in[p]}\frac{t_i^2\bar{\rho}_i(L_{i,j}^1)^2}{\underline{\rho}_i}\left(f^\star-f_j^\star\right)+\sum_{i=1}^{p}\frac{t_i^2\bar{\rho}_iL_{i,j}^0}{\underline{\rho}_i}\right)+\sum_{i=1}^{p}\frac{t_i^2L_i^0}{2} \\
&\quad +\sum_{i=1}^{p}\frac{2t_i^2\bar{\rho}_i\bar{L}_i^0}{\underline{\rho}_i\beta_i}+\sum_{i=1}^{p}\frac{2t_i^2\bar{\rho}_i\sqrt{1-\alpha_D}\bar{L}_i^0}{\underline{\rho}_i(1-\sqrt{1-\alpha_D})}+\sum_{i=1}^{p}\frac{4t_i\bar{\rho}_i\sqrt{1-\alpha_D}\beta_i\sigma_i}{1-\sqrt{1-\alpha_D}}+\sum_{i=1}^{p}2t_i\bar{\rho}_i\sigma_i\sqrt{\frac{\beta_i}{n}}.
\end{aligned}$$

Since $\delta^k \leq \Psi^k$, it follows that

$$\begin{aligned}
&\Psi^{k+1} \\
&\leq \left(1+2\max_{i\in[p]}(t_i^2(L_i^1)^2)+\frac{8\sqrt{1-\alpha_D}}{1-\sqrt{1-\alpha_D}}\max_{i\in[p],j\in[n]}\left(\frac{t_i^2\bar{\rho}_i}{\underline{\rho}_i}(L_{i,j}^1)^2\right)\right)\Psi^k \\
&\quad -\sum_{i=1}^{p}t_i\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right]+\sum_{i=1}^{p}2t_i\bar{\rho}_i(1-\beta_i)^kP_i^0 \\
&\quad +8\max_{i\in[p],j\in[n]}\left(\frac{t_i^2\bar{\rho}_i(L_{i,j}^1)^2}{\underline{\rho}_i}\right)\sum_{l=0}^{k-1}\max_{i\in[p]}((1-\beta_i)^{k-l}\Psi^l) \\
&\quad +8\frac{1}{n}\sum_{j=1}^{n}\left(\max_{i\in[p]}\frac{t_i^2\bar{\rho}_i(L_{i,j}^1)^2}{\underline{\rho}_i}\sum_{l=0}^{k-1}\max_{i\in[p]}\left((1-\beta_i)^{k-l}\right)\left(f^\star-f_j^\star\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&\quad +2\sum_{i=1}^{p}\frac{t_i^2\bar{\rho}_i\bar{L}_i^0}{\underline{\rho}_i}\sum_{l=0}^{k-1}(1-\beta_i)^{k-l}+\frac{8\sqrt{1-\alpha_D}}{1-\sqrt{1-\alpha_D}}\frac{1}{n}\sum_{j=1}^{n}\left(\max_{i\in[p]}\frac{t_i^2\bar{\rho}_i(L_{i,j}^1)^2}{\underline{\rho}_i}\left(f^\star-f_j^\star\right)\right)\\
&\quad +\frac{2\sqrt{1-\alpha_D}}{1-\sqrt{1-\alpha_D}}\sum_{i=1}^{p}\frac{t_i^2\bar{\rho}_i\bar{L}_i^0}{\underline{\rho}_i}+\sum_{i=1}^{p}t_i^2L_i^0\\
&\quad +\sum_{i=1}^{p}\frac{2t_i^2\bar{\rho}_i\bar{L}_i^0}{\underline{\rho}_i\beta_i}+\sum_{i=1}^{p}\frac{2t_i^2\bar{\rho}_i\sqrt{1-\alpha_D}\bar{L}_i^0}{\underline{\rho}_i(1-\sqrt{1-\alpha_D})}+\sum_{i=1}^{p}\frac{4t_i\bar{\rho}_i\sqrt{1-\alpha_D}\beta_i\sigma_i}{1-\sqrt{1-\alpha_D}}+\sum_{i=1}^{p}2t_i\bar{\rho}_i\sigma_i\sqrt{\frac{\beta_i}{n}}\\
\leq&\quad \left(1+\underbrace{2\max_{i\in[p]}(t_i^2(L_i^1)^2)+\frac{8\sqrt{1-\alpha_D}}{1-\sqrt{1-\alpha_D}}\max_{i\in[p],j\in[n]}\left(\frac{t_i^2\bar{\rho}_i(L_{i,j}^1)^2}{\underline{\rho}_i}\right)}_{:=C_1}\right)\Psi^k\\
&\quad -\sum_{i=1}^{p}t_i\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right]+\sum_{i=1}^{p}2t_i\bar{\rho}_i(1-\beta_i)^kP_i^0\\
&\quad +\underbrace{8\max_{i\in[p],j\in[n]}\left(\frac{t_i^2\bar{\rho}_i(L_{i,j}^1)^2}{\underline{\rho}_i}\right)}_{:=C_2}\sum_{l=0}^{k-1}((1-\beta_{\min})^{k-l}\Psi^l)\\
&\quad +\underbrace{8\left(\frac{1}{\beta_{\min}}+\frac{\sqrt{1-\alpha_D}}{1-\sqrt{1-\alpha_D}}\right)\frac{1}{n}\sum_{j=1}^{n}\left(\max_{i\in[p]}\frac{t_i^2\bar{\rho}_i(L_{i,j}^1)^2}{\underline{\rho}_i}\left(f^\star-f_j^\star\right)\right)}_{:=C_3}\\
&\quad +\sum_{i=1}^{p}t_i^2\underbrace{\left(L_i^0+\frac{4\bar{\rho}_i\bar{L}_i^0}{\underline{\rho}_i\beta_i}+\frac{4\bar{\rho}_i\sqrt{1-\alpha_D}\bar{L}_i^0}{\underline{\rho}_i(1-\sqrt{1-\alpha_D})}\right)}_{:=C_{4,i}}+\sum_{i=1}^{p}t_i\underbrace{\bar{\rho}_i\sigma_i\left(\frac{4\sqrt{1-\alpha_D}\beta_i}{1-\sqrt{1-\alpha_D}}+2\sqrt{\frac{\beta_i}{n}}\right)}_{:=C_{5,i}}\\
=&\quad (1+C_1)\,\Psi^k-\sum_{i=1}^{p}t_i\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right]+\sum_{i=1}^{p}2t_i\bar{\rho}_i(1-\beta_i)^kP_i^0\\
&\quad +C_2\sum_{l=0}^{k-1}((1-\beta_{\min})^{k-l}\Psi^l)+C_3+\sum_{i=1}^{p}t_i^2C_{4,i}+\sum_{i=1}^{p}t_iC_{5,i}.
\end{aligned}
$$

Now, define a weighting sequence $w^k := \frac{w^{k-1}}{1+C_1+\frac{C_2}{\beta_{\min}}}$, where $w^{-1}=1$. Then, multiplying the above inequality by w^k and summing over the first $K+1$ iterations, we obtain

$$
\begin{aligned}
&\sum_{k=0}^{K}w^k\Psi^{k+1}\\
\leq&\quad \sum_{k=0}^{K}w^k\left(1+C_1\right)\Psi^k+\sum_{k=0}^{K}w^kC_2\sum_{l=0}^{k-1}((1-\beta_{\min})^{k-l}\Psi^l)-\sum_{k=0}^{K}w^k\sum_{i=1}^{p}t_i\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right]\\
&\quad +\sum_{k=0}^{K}w^k\sum_{i=1}^{p}2t_i\bar{\rho}_i(1-\beta_i)^kP_i^0+\sum_{k=0}^{K}w^kC_3+\sum_{k=0}^{K}w^k\sum_{i=1}^{p}t_i^2C_{4,i}+\sum_{k=0}^{K}w^k\sum_{i=1}^{p}t_iC_{5,i}
\end{aligned}
$$

$$\begin{aligned}
&= (1+C_1)\sum_{k=0}^{K} w^k\Psi^k + C_2\sum_{k=0}^{K} w^k\sum_{l=0}^{k-1}((1-\beta_{\min})^{k-l}\Psi^l) - \sum_{k=0}^{K} w^k\sum_{i=1}^{p} t_i\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right] \\
&\quad + \sum_{k=0}^{K} w^k\sum_{i=1}^{p} 2t_i\bar{\rho}_i(1-\beta_i)^k P_i^0 + W^K C_3 + W^K\sum_{i=1}^{p} t_i^2 C_{4,i} + W^K\sum_{i=1}^{p} t_i C_{5,i}.
\end{aligned}$$

where $W^K := \sum_{k=0}^{K} w^k$. Since, by definition, $w^k \le w^{k-1} \le w^{-1} = 1$, we have

$$\begin{aligned}
&\sum_{k=0}^{K} w^k\Psi^{k+1} \\
&\le (1+C_1)\sum_{k=0}^{K} w^k\Psi^k + C_2\sum_{k=0}^{K}\sum_{l=0}^{k-1}(w^l(1-\beta_{\min})^{k-l}\Psi^l) - \sum_{k=0}^{K} w^k\sum_{i=1}^{p} t_i\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right] \\
&\quad + \sum_{k=0}^{K}\sum_{i=1}^{p} 2t_i\bar{\rho}_i(1-\beta_i)^k P_i^0 + W^K C_3 + W^K\sum_{i=1}^{p} t_i^2 C_{4,i} + W^K\sum_{i=1}^{p} t_i C_{5,i} \\
&\le (1+C_1)\sum_{k=0}^{K} w^k\Psi^k + C_2\sum_{l=0}^{\infty}(1-\beta_{\min})^l\sum_{k=0}^{K} w^k\Psi^k - \sum_{k=0}^{K} w^k\sum_{i=1}^{p} t_i\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right] \\
&\quad +2\sum_{i=1}^{p}\frac{t_i\bar{\rho}_i}{\beta_i}P_i^0 + W^K C_3 + W^K\sum_{i=1}^{p} t_i^2 C_{4,i} + W^K\sum_{i=1}^{p} t_i C_{5,i} \\
&= \left(1 + C_1 + \frac{C_2}{\beta_{\min}}\right)\sum_{k=0}^{K} w^k\Psi^k - \sum_{k=0}^{K} w^k\sum_{i=1}^{p} t_i\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right] \\
&\quad +2\sum_{i=1}^{p}\frac{t_i\bar{\rho}_i}{\beta_i}P_i^0 + W^K C_3 + W^K\sum_{i=1}^{p} t_i^2 C_{4,i} + W^K\sum_{i=1}^{p} t_i C_{5,i} \\
&= \sum_{k=0}^{K} w^{k-1}\Psi^k - \sum_{k=0}^{K} w^k\sum_{i=1}^{p} t_i\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right] + 2\sum_{i=1}^{p}\frac{t_i\bar{\rho}_i}{\beta_i}P_i^0 \\
&\quad +W^K C_3 + W^K\sum_{i=1}^{p} t_i^2 C_{4,i} + W^K\sum_{i=1}^{p} t_i C_{5,i}.
\end{aligned}$$

Rearranging the terms and dividing by W^K gives

$$\begin{aligned}
&\min_{k=0,\dots,K}\sum_{i=1}^{p} t_i\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right] \\
&\le \sum_{k=0}^{K}\sum_{i=1}^{p}\frac{w^k}{W^K}t_i\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right] \\
&\le \frac{1}{W^K}\sum_{k=0}^{K}\left(w^{k-1}\Psi^k - w^k\Psi^{k+1}\right) + \frac{2}{W^K}\sum_{i=1}^{p}\frac{t_i\bar{\rho}_i}{\beta_i}P_i^0 + C_3 + \sum_{i=1}^{p} t_i^2 C_{4,i} + \sum_{i=1}^{p} t_i C_{5,i} \\
&\le \frac{\Psi^0}{W^K} + \frac{2}{W^K}\sum_{i=1}^{p}\frac{t_i\bar{\rho}_i}{\beta_i}P_i^0 + C_3 + \sum_{i=1}^{p} t_i^2 C_{4,i} + \sum_{i=1}^{p} t_i C_{5,i}.
\end{aligned}$$

Now, note that

$$W^K = \sum_{k=0}^{K} w^k \geq (K+1)w^K = \frac{(K+1)w^{-1}}{(1 + C_1 + \frac{C_2}{\beta_{\min}})^{K+1}} \geq \frac{K+1}{\exp\left((K+1)(C_1 + \frac{C_2}{\beta_{\min}})\right)}.$$

Taking $t_i = \frac{\eta_i}{(K+1)^{3/4}}$, where $\eta_i^2 \leq \min\left\{\frac{(K+1)^{1/2}}{6(L_i^1)^2}, \frac{(1-\sqrt{1-\alpha_D})\underline{\rho}_i(K+1)^{1/2}}{24\sqrt{1-\alpha_D}\bar{\rho}_i(L_{i,\max}^1)^2}, \frac{\beta_{\min}\underline{\rho}_i(K+1)^{1/2}}{24\bar{\rho}_i(L_{i,\max}^1)^2}, 1\right\}$ to ensure that

$$\begin{aligned} 2(K+1)\max_{i\in[p]}(t_i^2(L_i^1)^2) &\leq \frac{1}{3}, \\ (K+1)\frac{8\sqrt{1-\alpha_D}}{1-\sqrt{1-\alpha_D}} \max_{i\in[p],j\in[n]} \left(\frac{t_i^2\bar{\rho}_i(L_{i,j}^1)^2}{\underline{\rho}_i}\right) &\leq \frac{1}{3}, \\ (K+1)\frac{8}{\beta_{\min}} \max_{i\in[p],j\in[n]} \left(\frac{t_i^2\bar{\rho}_i(L_{i,j}^1)^2}{\underline{\rho}_i}\right) &\leq \frac{1}{3}, \end{aligned}$$

we have $(K+1)(C_1 + \frac{C_2}{\beta_{\min}}) \leq 1$, and so $W^K \geq \frac{K+1}{\exp(1)} \geq \frac{K+1}{3}$. Therefore,

$$\begin{aligned} &\min_{k=0,\ldots,K} \sum_{i=1}^{p} t_i \mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right] \\ \leq\ & \frac{3\Psi^0}{K+1} + \frac{6}{K+1}\sum_{i=1}^{p}\frac{t_i\bar{\rho}_i}{\beta_i}P_i^0 + C_3 + \sum_{i=1}^{p} t_i^2 C_{4,i} + \sum_{i=1}^{p} t_i C_{5,i} \\ =\ & \frac{3\Psi^0}{K+1} + \frac{6}{K+1}\sum_{i=1}^{p}\frac{\eta_i\bar{\rho}_i}{\beta_i(K+1)^{3/4}}P_i^0 \\ & + \frac{8}{(K+1)^{3/2}}\left(\frac{1}{\beta_{\min}} + \frac{\sqrt{1-\alpha_D}}{1-\sqrt{1-\alpha_D}}\right)\frac{1}{n}\sum_{j=1}^{n}\left(\max_{i\in[p]}\frac{\eta_i^2\bar{\rho}_i(L_{i,j}^1)^2}{\underline{\rho}_i}\left(f^\star - f_j^\star\right)\right) \\ & + \sum_{i=1}^{p}\frac{\eta_i^2}{(K+1)^{3/2}}\left(L_i^0 + \frac{4\bar{\rho}_i\bar{L}_i^0}{\underline{\rho}_i\beta_i} + \frac{4\bar{\rho}_i\sqrt{1-\alpha_D}\bar{L}_i^0}{\underline{\rho}_i(1-\sqrt{1-\alpha_D})}\right) \\ & + \sum_{i=1}^{p}\frac{\eta_i}{(K+1)^{3/4}}\bar{\rho}_i\sigma_i\left(\frac{4\sqrt{1-\alpha_D}\beta_i}{1-\sqrt{1-\alpha_D}} + 2\sqrt{\frac{\beta_i}{n}}\right). \end{aligned}$$

Lastly, dividing by $\frac{1}{p}\sum_{l=1}^{p} t_l = \frac{1}{(K+1)^{3/4}}\frac{1}{p}\sum_{l=1}^{p}\eta_l$ gives

$$\begin{aligned} &\min_{k=0,\ldots,K} \sum_{i=1}^{p} \frac{\eta_i}{\frac{1}{p}\sum_{l=1}^{p}\eta_l} \mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right] \\ \leq\ & \frac{3\Psi^0}{(K+1)^{1/4}\frac{1}{p}\sum_{l=1}^{p}\eta_l} + \frac{6}{K+1}\sum_{i=1}^{p}\frac{\eta_i}{\frac{1}{p}\sum_{l=1}^{p}\eta_l}\frac{\bar{\rho}_i}{\beta_i}P_i^0 \\ & + \frac{8}{(K+1)^{3/4}}\left(\frac{1}{\beta_{\min}} + \frac{\sqrt{1-\alpha_D}}{1-\sqrt{1-\alpha_D}}\right)\frac{1}{n}\sum_{j=1}^{n}\frac{\max_{i\in[p]}\frac{\eta_i^2\bar{\rho}_i(L_{i,j}^1)^2}{\underline{\rho}_i}}{\frac{1}{p}\sum_{l=1}^{p}\eta_l}\left(f^\star - f_j^\star\right) \end{aligned}$$

$$
\begin{aligned}
&+\sum_{i=1}^{p}\frac{\eta_i^2}{(K+1)^{3/4}\frac{1}{p}\sum_{l=1}^{p}\eta_l}\left(L_i^0+\frac{4\bar{\rho}_i\bar{L}_i^0}{\underline{\rho}_i\beta_i}+\frac{4\bar{\rho}_i\sqrt{1-\alpha_D}\bar{L}_i^0}{\underline{\rho}_i(1-\sqrt{1-\alpha_D})}\right)\\
&+\sum_{i=1}^{p}\frac{\eta_i\bar{\rho}_i\sigma_i}{\frac{1}{p}\sum_{l=1}^{p}\eta_l}\left(\frac{4\sqrt{1-\alpha_D}\beta_i}{1-\sqrt{1-\alpha_D}}+2\sqrt{\frac{\beta_i}{n}}\right)\\
= \quad &\frac{3\Psi^0}{(K+1)^{1/4}\frac{1}{p}\sum_{l=1}^{p}\eta_l}+\frac{6}{(K+1)^{1/2}}\sum_{i=1}^{p}\frac{\eta_i\bar{\rho}_i}{\frac{1}{p}\sum_{l=1}^{p}\eta_l}P_i^0\\
&+\left(\frac{8}{(K+1)^{1/4}}+\frac{8\sqrt{1-\alpha_D}}{(1-\sqrt{1-\alpha_D})(K+1)^{3/4}}\right)\frac{1}{n}\sum_{j=1}^{n}\frac{\max_{i\in[p]}\frac{\eta_i^2\bar{\rho}_i(L_{i,j}^1)^2}{\underline{\rho}_i}}{\frac{1}{p}\sum_{l=1}^{p}\eta_l}\left(f^\star-f_j^\star\right)\\
&+\sum_{i=1}^{p}\frac{\eta_i^2}{\frac{1}{p}\sum_{l=1}^{p}\eta_l}\left(\frac{L_i^0}{(K+1)^{3/4}}+\frac{4\bar{\rho}_i\bar{L}_i^0}{\underline{\rho}_i(K+1)^{1/4}}+\frac{4\bar{\rho}_i\sqrt{1-\alpha_D}\bar{L}_i^0}{\underline{\rho}_i(1-\sqrt{1-\alpha_D})(K+1)^{3/4}}\right)\\
&+\sum_{i=1}^{p}\frac{\eta_i\bar{\rho}_i\sigma_i}{\frac{1}{p}\sum_{l=1}^{p}\eta_l}\left(\frac{4\sqrt{1-\alpha_D}}{(1-\sqrt{1-\alpha_D})(K+1)^{1/2}}+\frac{2}{\sqrt{n}(K+1)^{1/4}}\right),
\end{aligned}
$$

where in the last equality we set $\beta = \frac{1}{(K+1)^{1/2}}$. ■

Proof [Proof of Theorem 37] Substituting the initialization, we have

$$
\begin{aligned}
P_i^0 \quad &:= \quad \mathbb{E}\left[\left\|\nabla_i f(X^0)-M_i^0\right\|_2\right]=\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\left(\nabla_i f_j(X^0)-\nabla_i f_j(X^0;\xi_j^0)\right)\right\|_2\right]\\
&\leq \quad \sqrt{\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\left(\nabla_i f_j(X^0)-\nabla_i f_j(X^0;\xi_j^0)\right)\right\|_2^2\right]}\overset{(10)}{\leq}\frac{\sigma_i}{\sqrt{n}},\\
\tilde{P}_i^0 \quad &:= \quad \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|\nabla_i f_j(X^0)-M_{i,j}^0\right\|_2\right]=\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|\nabla_i f_j(X^0)-\nabla_i f_j(X^0;\xi_j^0)\right\|_2\right]\leq\sigma_i,\\
\tilde{S}_i^0 \quad &:= \quad \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|M_{i,j}^0-G_{i,j}^0\right\|_2\right]=\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|\nabla_i f_j(X^0;\xi_j^0)-\mathcal{C}_{i,j}^0(\nabla_i f_j(X^0;\xi_j^0))\right\|_2\right]\\
&\overset{(1)}{\leq} \quad \sqrt{1-\alpha_D}\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|\nabla_i f_j(X^0;\xi_j^0)\right\|_2\right]\\
&\leq \quad \sqrt{1-\alpha_D}\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|\nabla_i f_j(X^0;\xi_j^0)-\nabla_i f_j(X^0)\right\|_2\right]\overset{(10)}{\leq}\sqrt{1-\alpha_D}\sigma_i,
\end{aligned}
$$

and hence

$$
\begin{aligned}
\Psi^0 \quad := \quad & f(X^0)-f^\star+\sum_{i=1}^{p}\frac{2t_i\bar{\rho}_i}{1-\sqrt{1-\alpha_D}}\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|M_{i,j}^0-G_{i,j}^0\right\|_2\right]\\
&+\sum_{i=1}^{p}\frac{2t_i\bar{\rho}_i\sqrt{1-\alpha_D}}{1-\sqrt{1-\alpha_D}}\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\left\|\nabla_i f_j(X^0)-M_{i,j}^0\right\|_2\right]
\end{aligned}
$$

$$
\begin{aligned}
&\le\quad f(X^0) - f^\star + \sum_{i=1}^{p} \frac{2t_i\bar{\rho}_i}{1-\sqrt{1-\alpha_D}}\sqrt{1-\alpha_D}\sigma_i + \sum_{i=1}^{p}\frac{2t_i\bar{\rho}_i\sqrt{1-\alpha_D}}{1-\sqrt{1-\alpha_D}}\sigma_i \\
&=\quad f(X^0) - f^\star + \sum_{i=1}^{p}\frac{4\sqrt{1-\alpha_D}t_i\bar{\rho}_i\sigma_i}{1-\sqrt{1-\alpha_D}}.
\end{aligned}
$$

Substituting this in the rate, we get

$$
\begin{aligned}
&\min_{k=0,\ldots,K}\sum_{i=1}^{p}\frac{\eta_i}{\frac{1}{p}\sum_{l=1}^{p}\eta_l}\mathbb{E}\left[\left\|\nabla_i f(X^k)\right\|_{(i)\star}\right] \\
&\le\quad \frac{3}{(K+1)^{1/4}\frac{1}{p}\sum_{l=1}^{p}\eta_l}\left(f(X^0)-f^\star+\sum_{i=1}^{p}\frac{4\sqrt{1-\alpha_D}\eta_i\bar{\rho}_i\sigma_i}{(K+1)^{3/4}(1-\sqrt{1-\alpha_D})}\right) \\
&\qquad + \frac{6}{(K+1)^{1/2}}\sum_{i=1}^{p}\frac{\bar{\rho}_i\eta_i\sigma_i}{\sqrt{n}\frac{1}{p}\sum_{l=1}^{p}\eta_l} \\
&\qquad + \left(\frac{8}{(K+1)^{1/4}}+\frac{8\sqrt{1-\alpha_D}}{(K+1)^{3/4}(1-\sqrt{1-\alpha_D})}\right)\frac{1}{n}\sum_{j=1}^{n}\frac{\max_{i\in[p]}\eta_i^2\frac{\bar{\rho}_i}{\underline{\rho}_i}(L_{i,j}^1)^2}{\frac{1}{p}\sum_{l=1}^{p}\eta_l}\left(f^\star-f_j^\star\right) \\
&\qquad + \sum_{i=1}^{p}\frac{\eta_i^2}{\frac{1}{p}\sum_{l=1}^{p}\eta_l}\left(\frac{L_i^0}{(K+1)^{3/4}}+\frac{4\bar{\rho}_i\bar{L}_i^0}{\underline{\rho}_i(K+1)^{1/4}}+\frac{4\bar{\rho}_i\sqrt{1-\alpha_D}\bar{L}_i^0}{\underline{\rho}_i(K+1)^{3/4}(1-\sqrt{1-\alpha_D})}\right) \\
&\qquad + \sum_{i=1}^{p}\frac{\eta_i\bar{\rho}_i\sigma_i}{\frac{1}{p}\sum_{l=1}^{p}\eta_l}\left(\frac{4\sqrt{1-\alpha_D}}{(K+1)^{1/2}(1-\sqrt{1-\alpha_D})}+\frac{2}{\sqrt{n}(K+1)^{1/4}}\right).
\end{aligned}
$$

■

Appendix G. Experiments

This section provides additional experimental results and setup details complementing Section 5.

G.1. Setup details

Tables 2 to 4 summarize the model and optimizer hyperparameters. The *scale* parameters (Hidden/-Head Scale) in Table 4 specify the LMO trust-region radius as

$$\text{radius} = \text{scale} \times \text{learning rate},$$

following Pethick et al. [54], Riabinin et al. [61].

Table 2: `NanoGPT`-124M model configuration.

Hyperparameter	**Value**
Total Parameters	124M
Vocabulary Size	50,304
Number of Transformer Layers	12
Attention Heads	6
Hidden Size	768
FFN Hidden Size	3,072
Positional Embedding	RoPE [71]
Activation Function	Squared ReLU [68]
Normalization	RMSNorm [86]
Bias Parameters	None

Table 3: `MediumGPT`-335M model configuration.

Hyperparameter	**Value**
Total Parameters	335M
Vocabulary Size	50,304
Number of Transformer Layers	24
Attention Heads	16
Hidden Size	1024
FFN Hidden Size	4096
Positional Embedding	RoPE [71]
Activation Function	Squared ReLU [68]
Normalization	RMSNorm [86]
Bias Parameters	None

G.2. Top$K$ compression details

Top$K$ compressor requires transmitting both the selected values and their corresponding indices to reconstruct the original tensors. At high compression levels, this introduces significant communication overhead, especially in compositional schemes such as Top$K$ combined with the Natural

Table 4: Optimizer configuration.

Hyperparameter	**Value**
Sequence Length	1024
Batch Size	256
Optimizer	EF21-Muon
Weight Decay	0
Hidden Layer Norm	Spectral norm
Hidden Layer Scale	50
Newton–Schulz Iterations	5
Embedding and Head Layers Norm	ℓ_∞ norm
Embedding and Head Layers Scale	3000
Initial Learning Rate	For non-compressed: 3.6×10^{-4}
Learning Rate Schedule	Constant followed by linear decreasing
Learning Rate Constant Phase Length	40% of tokens
Momentum	0.9

compressor, where the cost of transmitting indices can even exceed that of the quantized values. To illustrate this effect, we analyze the largest parameter matrices in the NanoGPT model: the token embedding layer and the classification head, each of size $50,304 \times 768$. Representing an index for any element in these matrices requires $\log_2(50{,}304 \cdot 768) < 26$ bits. We use this calculation when visualizing communication costs.

G.3. Learning rate ablation

To ensure a fair and robust comparison, we perform a learning rate hyperparameter sweep for each compression configuration, as detailed in Figure 3. For every method, the search space is initialized at the optimal learning rate of the uncompressed baseline (taken from the Gluon repository [60]) and spans downward by up to an order of magnitude. We consistently observe that more aggressive compression schemes require a smaller learning rate for stable convergence.

This tuning protocol is applied uniformly across all experiments for models trained with 2.5B (Section G.5) and 5B token budgets.

G.4. Compression level ablation

This section presents an ablation study on the compression ratio, governed by the parameter K. Figures 4 and 5 illustrate the convergence curves for various compression configurations, each trained with its optimal learning rate (see Section G.3). Figure 6 summarizes the final loss as a function of K.

Our results show that for Top$K$ and Rand$K$ compressors, an aggressive compression ratio of $K = 5\%$ quite severely impairs convergence (see Figure 6), while configurations with $K \geq 10\%$ achieve satisfactory loss reduction. When these compressors are composed with the Natural compressor, convergence degradation is more pronounced for $K = 10\%$ than for the less aggressive $K = 15\%$ setup.

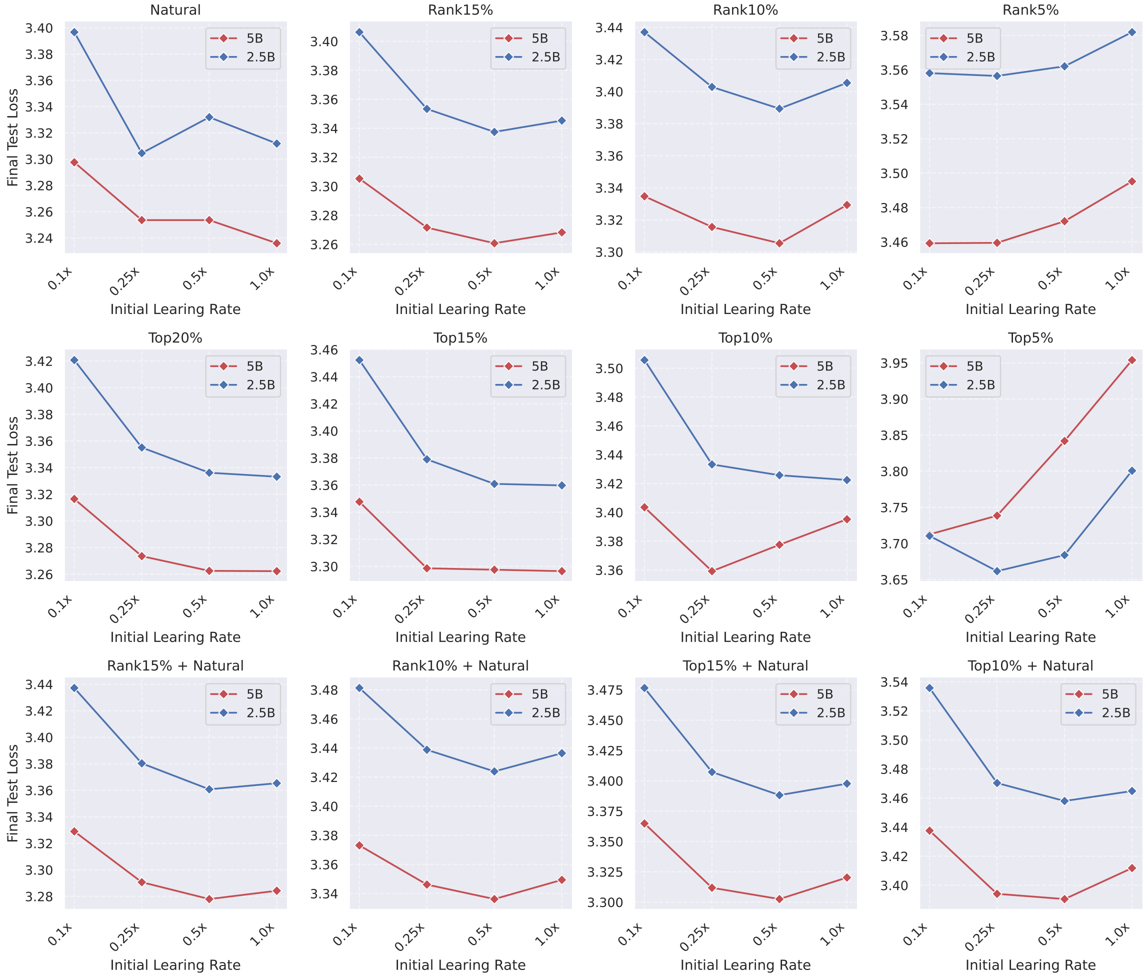

Figure 3: **Learning rate ablation.** The grid spans from the optimal learning rate of the non-compressed baseline, 3.6×10^{-4} (denoted as $1.0\times$), down to $0.1\times$. Red curves correspond to experiments processing 5B tokens (Section 5), while blue curves correspond to 2.5B tokens (Section G.5).

We also examine a more challenging loss threshold of 3.28 (Figure 7). The communication cost improvement at this threshold is even more pronounced than for 3.31 (Figure 1), but this comes at a cost: only a subset of compressors can reach the threshold within the 5B token budget.

G.5. 2.5B tokens experiment

In Section 5, we report runs with a 5B token budget ($> 40\times$ model size). Testing convergence over a large number of tokens is important, as the limitations of compressors relative to the baseline become more pronounced after many steps. At the same time, evaluating compressed runs with a smaller token budget is useful for cases with limited resources. We provide a learning rate ablation in Figure 3, a summarized comparison in Figure 6, and convergence trajectories for the 2.5B-token setup in Figures 8 and 9.

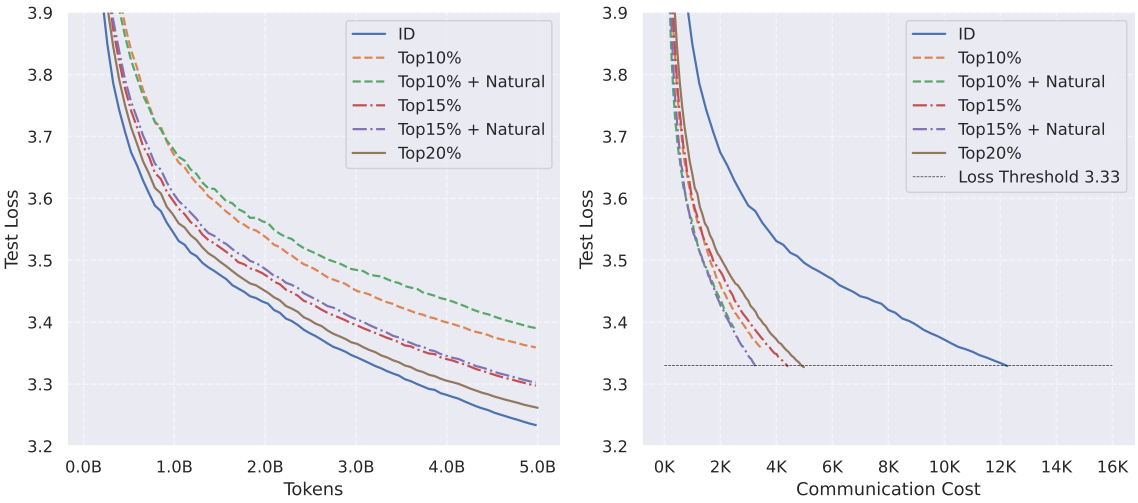

Figure 4: Left: **Test loss vs. $\#$ of tokens processed.** Right: **Test loss vs. $\#$ of bytes sent to the server from each worker** normalized by model size to reach test loss 3.33. Top$X\%$ = Top$K$ compressor with sparsification level $X\%$; ID = no compression.

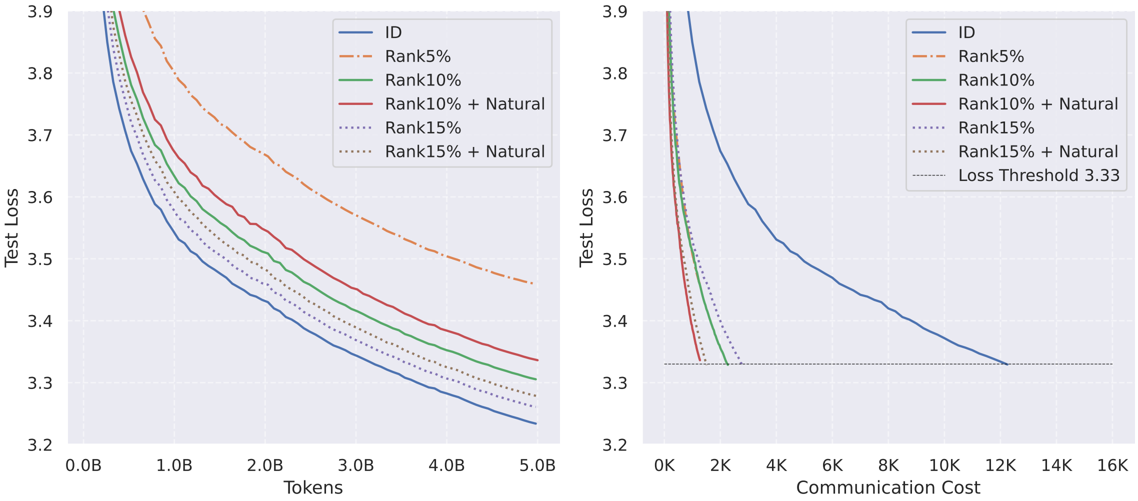

Figure 5: Left: **Test loss vs. $\#$ of tokens processed.** Right: **Test loss vs. $\#$ of bytes sent to the server from each worker** normalized by model size to reach test loss 3.33. Rank$X\%$ = Rank$K$ compressor with sparsification level $X\%$; ID = no compression.

G.6. MediumGPT experiment

To assess whether the patterns observed on NanoGPT scale to larger models, we conduct experiments on MediumGPT (335M parameters) [27] with 2.5B token budget. The model configuration is provided in Table 3. We compare the uncompressed baseline to EF21-Muon with the Natural compressor and evaluate convergence in terms of both tokens and bytes communicated.

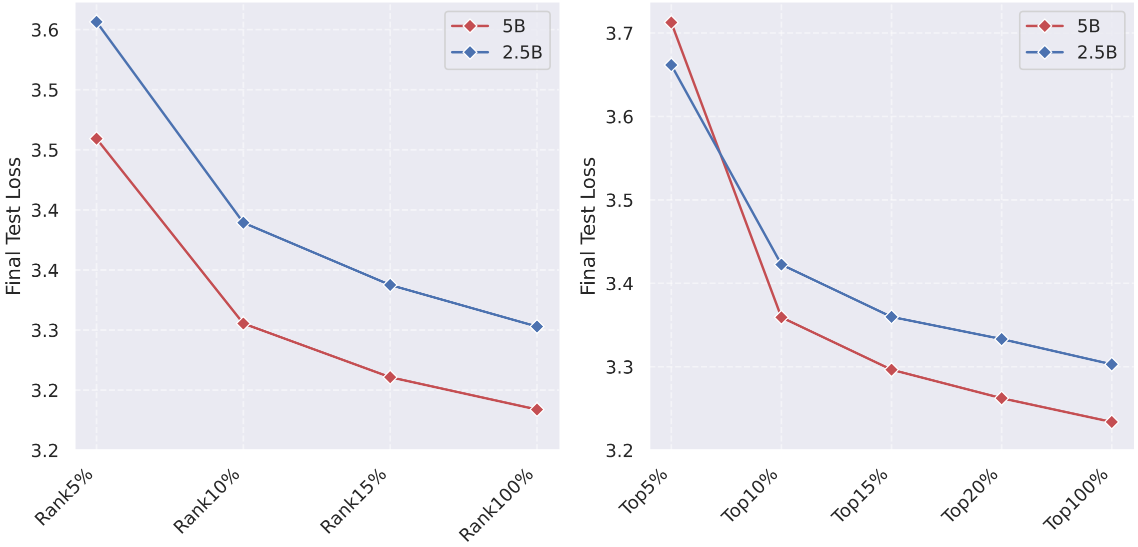

Figure 6: **Final test loss vs. compression parameter K.** Results are shown after processing 5B tokens (red) and 2.5B tokens (blue) for Rank$K$ (left) and Top$K$ (right) compressors. $K = 100\%$ corresponds to the non-compressed baseline. In the Top$K$ plot, the 2.5B setup outperforms 5B due to differences in scheduler behavior, as the runs execute a different number of steps.

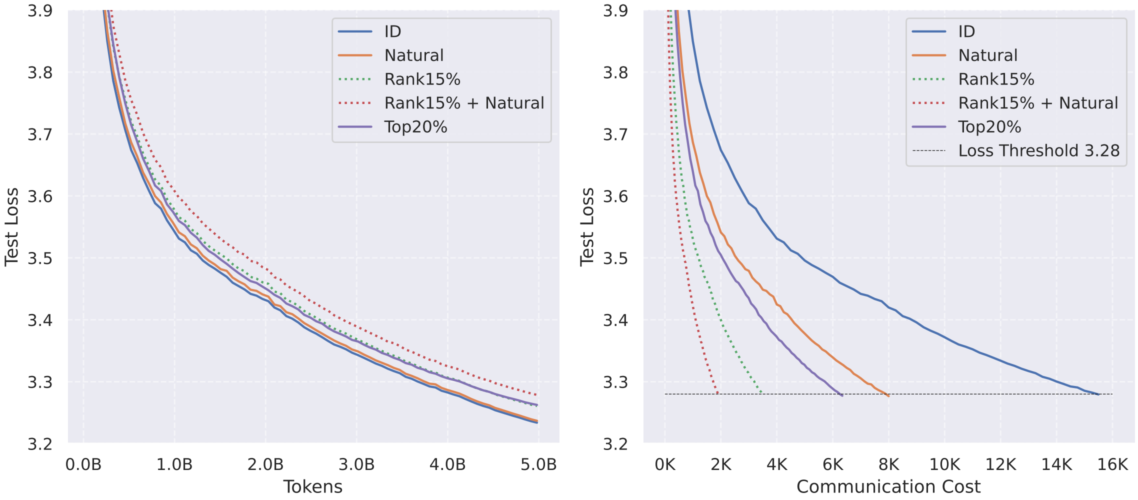

Figure 7: Left: **Test loss vs. # of tokens processed.** Right: **Test loss vs. # of bytes sent to the server from each worker** normalized by model size to reach test loss 3.28. Rank$X\%$/Top$X\%$ = Rank$K$/Top$K$ compressor with sparsification level $X\%$; ID = no compression.

We adopt the learning rate obtained from the sweep described in Section G.3 and use the same optimization and training setup as in the `NanoGPT` experiments. The LMO step scaling mechanism [53] is applied to ensure adaptivity across weight matrices of varying sizes.

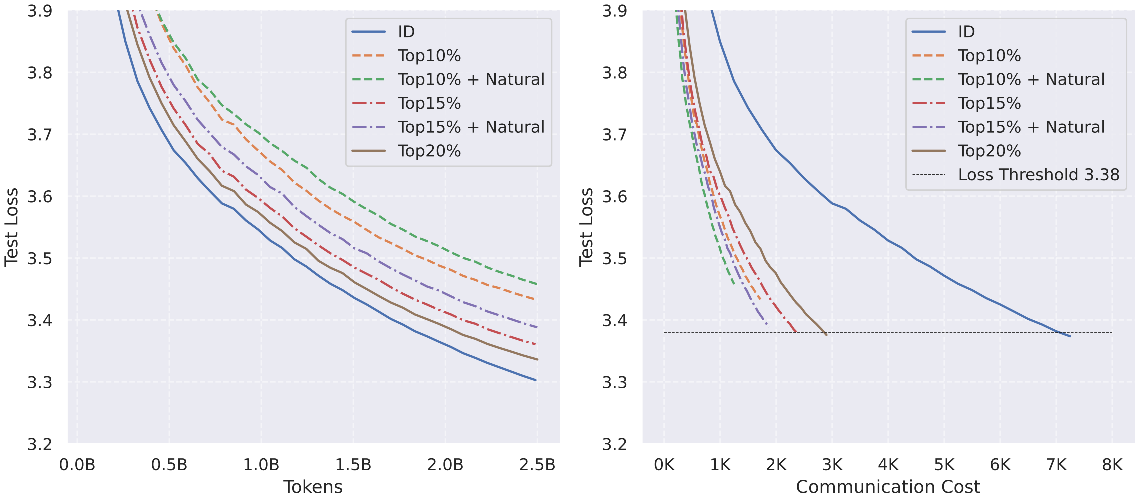

Figure 8: Left: **Test loss vs. # of tokens processed.** Right: **Test loss vs. # of bytes sent to the server from each worker** normalized by model size to reach test loss 3.38. Top$X\%$ = Top$K$ compressor with sparsification level $X\%$; ID = no compression. "+ Natural" corresponds to applying Natural compression after Top$K$ compressor.

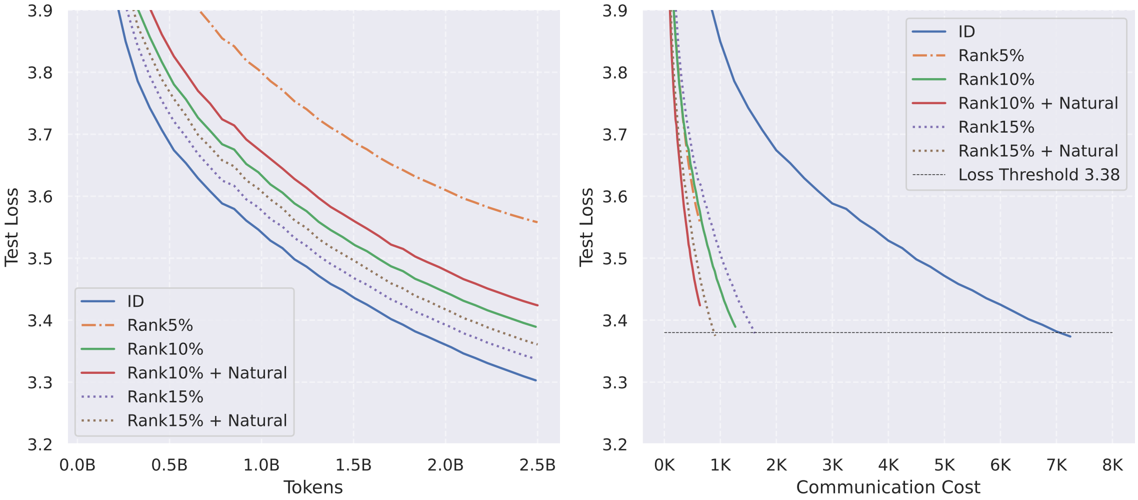

Figure 9: Left: **Test loss vs. # of tokens processed.** Right: **Test loss vs. # of bytes sent to the server from each worker** normalized by model size to reach test loss 3.38. Rank$X\%$ = Rank$K$ compressor with sparsification level $X\%$; ID = no compression. "+ Natural" corresponds to applying Natural compression after Rank$K$ compressor.

The resulting convergence curves are shown in Figure 10. We observe qualitatively similar behavior to the `NanoGPT` setting: Natural compression achieves close-to-baseline loss while substantially reducing communication cost.

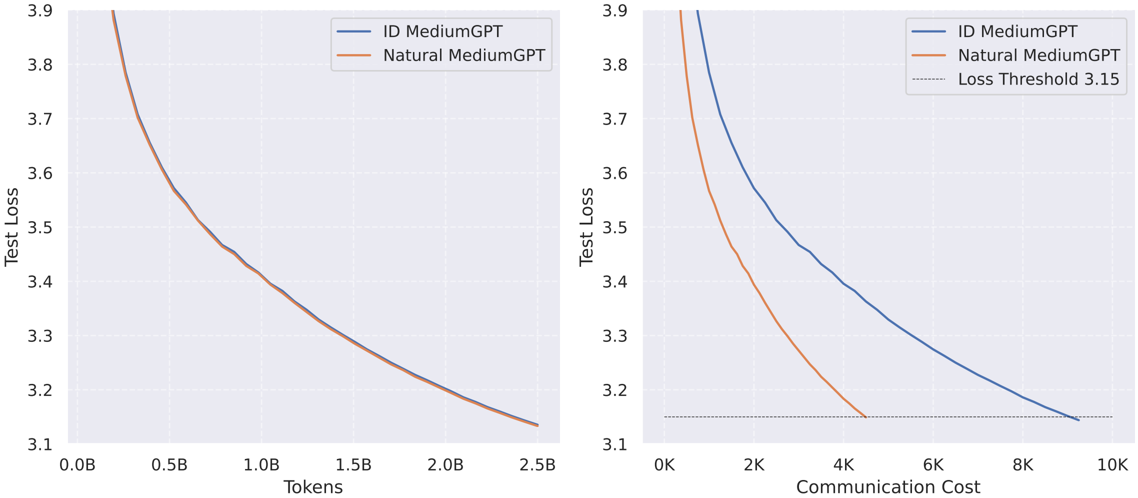

Figure 10: Left: **Test loss vs. $\#$ of tokens processed.** Right: **Test loss vs. $\#$ of bytes sent to the server from each worker** normalized by model size to reach test loss 3.15. ID = no compression.

G.7. Bidirectional compression

To complement the unidirectional compression experiments, we evaluate EF21-Muon in a fully bidirectional setup in which both server-to-worker and worker-to-server communication are compressed. We apply the Natural compressor in both directions. The training task matches the setup in Section G.5 (`NanoGPT` with a 2.5B token budget).

We follow the same hyperparameter selection protocol described in Section G.3. The corresponding learning rate sweep is shown in Figure 11. After tuning, we find that EF21-Muon remains effective in this more challenging bidirectional configuration, improving communication efficiency by approximately $2\times$ relative to the uncompressed baseline while achieving comparable convergence, as shown in Figure 12.

G.8. Limitations

Reporting results for all compressors on the same token budget (for instance, 5B) and then measuring the prefix needed to reach a given loss threshold may not be fully consistent, as results can be affected by the scheduler. To mitigate this, we use a relatively strong loss threshold that ensures a significant number of tokens are processed beyond the constant learning rate phase. Additionally, tuning the initial learning rate can help stabilize the results.

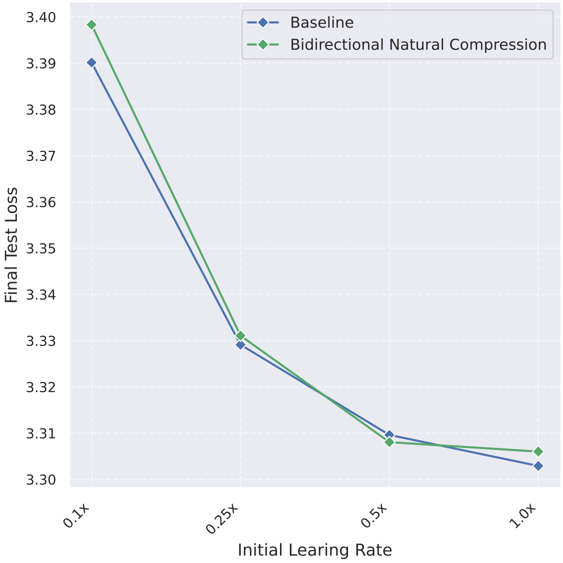

Figure 11: **Learning rate ablation for the bidirectional setup.** The grid spans from the optimal learning rate of the non-compressed baseline, 3.6×10^{-4} (denoted as $1.0\times$), down to $0.1\times$.

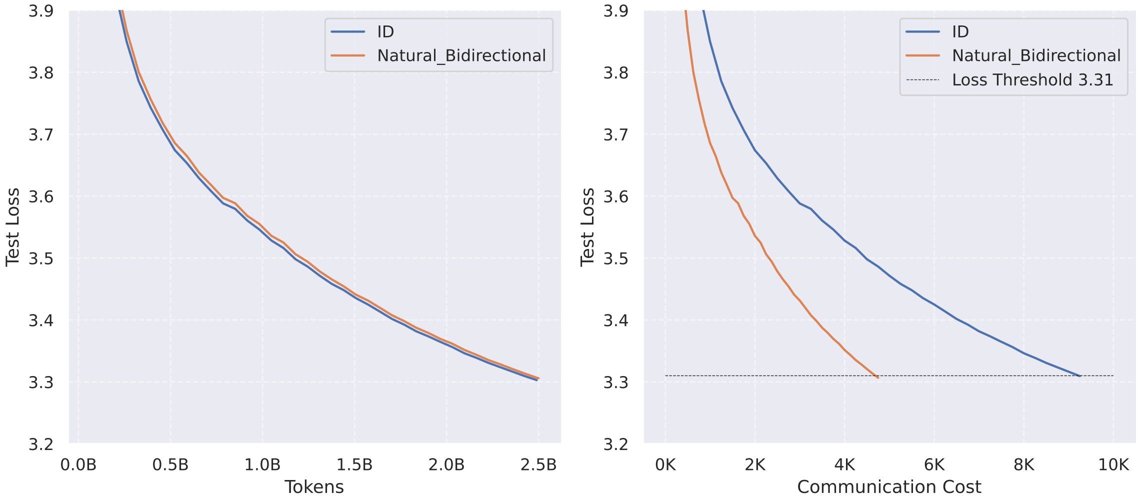

Figure 12: Left: **Test loss vs. # of tokens processed.** Right: **Test loss vs. # of bytes sent** normalized by model size to reach test loss 3.31. ID = no compression. Both s2w and w2s directions are compressed using the Natural compressor.

Appendix H. Useful Facts and Lemmas

For all $X, Y \in \mathcal{S}$, $Z \in \mathcal{S}^\star$ (where $\mathcal{S}^\star$ is the dual space of $\mathcal{S}$), $t > 0$ and $\alpha \in (0, 1]$, we have:

$$\|X + Y\|^2 \leq (1 + s)\|X\|^2 + (1 + s^{-1})\|Y\|^2, \tag{27}$$

$$\langle X, Z \rangle \leq \frac{\|X\|^2}{2s} + \frac{s\|Z\|_\star^2}{2}, \tag{28}$$

$$(1-\alpha)\left(1+\frac{\alpha}{2}\right) \leq 1-\frac{\alpha}{2}, \tag{29}$$

$$(1-\alpha)\left(1+\frac{2}{\alpha}\right) \leq \frac{2}{\alpha}, \tag{30}$$

$$\left\langle G, \mathrm{LMO}_{\mathcal{B}(X,t)}(G)\right\rangle = -t\,\|G\|_\star \tag{31}$$

$$\left\langle X, X^\sharp\right\rangle = \left\|X^\sharp\right\|^2, \tag{32}$$

$$\|X\|_\star = \left\|X^\sharp\right\|. \tag{33}$$

Lemma 39 (Riabinin et al. [61], Lemma 3) *Suppose that $x_1,\ldots,x_p,y_1,\ldots,y_p \in \mathbb{R}$, $\max_{i\in[p]}|x_i| > 0$ and $z_1,\ldots,z_p > 0$. Then*

$$\sum_{i=1}^{p}\frac{y_i^2}{z_i} \geq \frac{\left(\sum_{i=1}^p x_i y_i\right)^2}{\sum_{i=1}^p z_i x_i^2}.$$

Lemma 40 (Variance decomposition) *For any random vector $X \in \mathcal{S}$ and any non-random $c \in \mathcal{S}$, we have*

$$\mathbb{E}\left[\|X-c\|_2^2\right] = \mathbb{E}\left[\|X-\mathbb{E}[X]\|_2^2\right] + \|\mathbb{E}[X]-c\|_2^2.$$

Lemma 41 (Riabinin et al. [61], Lemma 1) *Let Assumption 8 hold. Then, for any $X, Y \in \mathcal{S}$,*

$$|f(Y)-f(X)-\langle\nabla f(X), Y-X\rangle| \leq \sum_{i=1}^{p}\frac{L_i^0 + L_i^1\|\nabla_i f(X)\|_{(i)\star}}{2}\|X_i - Y_i\|_{(i)}^2.$$

Lemma 42 *Let $\{A^k\}_{k\geq 0}$, $\{B_i^k\}_{k\geq 0}$, $i\in[p]$ be non-negative sequences such that*

$$A^{k+1} \leq (1+a_1)A^k - \sum_{i=1}^p B_i^k + a_2,$$

where $a_1, a_2 \geq 0$. Then

$$\min_{k=0,\ldots,K}\sum_{i=1}^p B_i^k \leq \frac{\exp(a_1(K+1))}{(K+1)}A^0 + a_2.$$

Proof Let us define a weighting sequence $w^k := \frac{w^{k-1}}{1+a_1}$, where $w^{-1} = 1$. Then

$$w^k A^{k+1} \leq w^k(1+a_1)A^k - w^k\sum_{i=1}^p B_i^k + w^k a_2 = w^{k-1}A^k - w^k\sum_{i=1}^p B_i^k + w^k a_2,$$

and hence

$$\min_{k=0,\ldots,K}\sum_{i=1}^p B_i^k \quad\leq\quad \frac{1}{\sum_{k=0}^K w^k}\sum_{k=0}^K w^k\sum_{i=1}^p B_i^k$$

$$
\begin{aligned}
&\leq \frac{1}{\sum_{k=0}^{K} w^k} \sum_{k=0}^{K} \left(w^{k-1} A^k - w^k A^{k+1}\right) + \frac{1}{\sum_{k=0}^{K} w^k} \sum_{k=0}^{K} w^k a_2 \\
&= \frac{1}{\sum_{k=0}^{K} w^k} \left(w^{-1} A^0 - w^K A^{K+1}\right) + a_2.
\end{aligned}
$$

Using the fact that $w^{-1} = 1$ and $\sum_{k=0}^{K} w^k = \sum_{k=0}^{K} \frac{1}{(1+a_1)^{k+1}} \geq \frac{K+1}{(1+a_1)^{K+1}}$, we get

$$
\min_{k=0,\ldots,K} \sum_{i=1}^{p} B_i^k \leq \frac{(1+a_1)^{K+1}}{(K+1)} \left(A^0 - w^K A^{K+1}\right) + a_2 \leq \frac{\exp(a_1(K+1))}{(K+1)} A^0 + a_2,
$$

which finishes the proof. ■