Scalable Distributed Stochastic Optimization via Bidirectional Compression: Beyond Pessimistic Limits

Grigory Begunov
AXXX[1], MSU[2]

Alexander Tyurin
AXXX[1], Applied AI Institute[3]

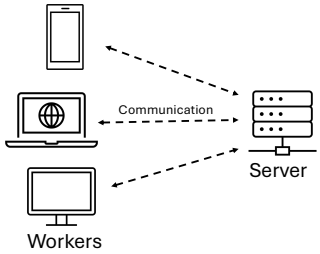
Abstract

In centralized, distributed, and federated learning with stochastic gradients and n workers, it was recently shown that it is infeasible to find an ε–stationary point faster than $\tilde{\Omega}(\min\{{}^{d\kappa L\Delta}/_{\varepsilon} + {}^{hL\Delta}/_{\varepsilon} + {}^{h\sigma^2 L\Delta}/_{n\varepsilon^2}, {}^{h\sigma^2 L\Delta}/_{\varepsilon^2} + {}^{hL\Delta}/_{\varepsilon}\})$ seconds in both homogeneous and heterogeneous settings under standard assumptions: L–smoothness, σ^2-bounded unbiased stochastic gradients, and lower boundedness of the function, i.e., $f(x) \geq f^*$ for all $x \in \mathbb{R}^d$, where $\Delta = f(x^0) - f^*$, h is the computation time, κ is the communication speed between the workers and the server, and d is the dimension of the iterates and gradients. This result is pessimistic since it does not allow a complexity in which both ${}^{d\kappa L\Delta}/_{\varepsilon}$ and ${}^{h\sigma^2 L\Delta}/_{\varepsilon^2}$ improve with n, even when using random sparsification techniques; moreover, this lower bound can be matched by either non-distributed SGD or vanilla Synchronous SGD, which reduces the impact of recent progress in the design of compression-based methods. In this work, we challenge this limitation and propose new compressed methods, Inkheart SGD and M4, and show that under an additional structural assumption, which is necessary due to the lower bound and which does not restrict the class of considered problems, we achieve new state-of-the-art time complexities that break this pessimistic barrier and allow scaling with the number of workers n.

1 Introduction

We consider a stochastic nonconvex optimization problem

$$\min_{x \in \mathbb{R}^d} f(x), \tag{1}$$



where $f : \mathbb{R}^d \to \mathbb{R}$ and d is the dimension of f. We assume that d is huge, which is indeed the case in modern machine learning and large language model training (Brown et al., 2020; Touvron et al., 2023). This problem is solved in the classical federated learning setup, where n workers, such as CPUs, GPUs, servers, or mobile devices, are connected to a central server via a communication channel (Konečný et al., 2016; McMahan et al., 2017). The workers compute stochastic gradients in parallel and send them to the server, which aggregates the information and communicates it back to the workers. We start by considering the homogeneous setup, but we also consider the heterogeneous setup in Section 3.

In the nonconvex setting, the objective is to obtain an ε-stationary point, that is, a (random) point $\bar{x} \in \mathbb{R}^d$ satisfying $\mathbb{E}[\|\nabla f(\bar{x})\|^2] \leq \varepsilon$ under the following standard optimization assumptions:

[1]AXXX, Moscow, Russia
[2]Lomonosov Moscow State University, Moscow, Russia
[3]Applied AI Institute, Moscow, Russia

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Assumption 1.1. Function f is L–smooth:

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\| \qquad \forall x, y \in \mathbb{R}^d.$$

Assumption 1.2. There exists $f^* \in \mathbb{R}$ such that $f(x) \geq f^*$ for all $x \in \mathbb{R}^d$. We define $\Delta := f(x^0) - f^*$, where x^0 is the starting point of optimization methods.

Assumption 1.3. The workers have access to unbiased stochastic gradients $\nabla f(x;\xi)$ with σ^2–bounded variance: $\mathbb{E}_\xi [\nabla f(x;\xi)] = \nabla f(x)$ and $\mathbb{E}_\xi[\|\nabla f(x;\xi) - \nabla f(x)\|^2] \leq \sigma^2$ for all $x \in \mathbb{R}^d$, where $\sigma > 0$ is some constant.

In order to compare methods, explain the main goal, and present the new theoretical results we consider the following assumption about the optimization environment. Note that the newly presented methods are valid and converge even without it.

Assumption 1.4. Each of the n workers requires at most h seconds to compute a stochastic gradient, and communication *from the server to any worker* (s2w communication) takes at most κ seconds per coordinate, and communication *from any worker to the server* (w2s communication) takes at most τ seconds per coordinate.

We consider settings with bidirectional communication costs, where transmitting information in either direction requires time. Under Assumption 1.4, sending a vector $v \in \mathbb{R}^d$ from the server to any worker takes $d \times \kappa$ seconds, while communication from any worker to the server requires $d \times \tau$ seconds. In contrast, many existing works, especially in the early development of federated learning algorithms, assume that communication *from the server to the workers* is free, i.e., $\kappa = 0$, which is arguably not realistic in practice, since communication over the Internet or 4G/5G networks can be costly in both directions (Huang et al., 2012; Narayanan et al., 2021).

1.1 Related work

Synchronous method. The centralized setting under the computation and communication costs assumptions is well studied. The standard baseline is the Synchronous SGD method: $x^{k+1} = x^k - \frac{\gamma}{n}\sum_{i=1}^n \nabla f(x^k;\xi_i^k)$, where γ is the step size and every worker computes one stochastic gradient (takes h seconds), sends it to the server (takes $d\tau$ seconds), which updates the point and sends it back to the workers (takes $d\kappa$ seconds). One can easily show that the iteration complexity of this method with a proper γ is $\mathcal{O}(L\Delta/\varepsilon + \sigma^2 L\Delta/n\varepsilon^2)$ (Lan, 2020) and the time complexity is $\mathcal{O}\left((h + d\tau + d\kappa)\left(L\Delta/\varepsilon + \sigma^2 L\Delta/n\varepsilon^2\right)\right)$. Using the minibatching technique, i.e., calculating $\sum_{j=1}^b \nabla f(x^k;\xi_{ij}^k)$ instead of $\nabla f(x^k;\xi_i^k)$ with a proper b, this complexity can be improved to:

$$\mathcal{O}\left((d\kappa + d\tau)\frac{L\Delta}{\varepsilon} + h\frac{L\Delta}{\varepsilon} + h\frac{\sigma^2 L\Delta}{n\varepsilon^2}\right), \tag{2}$$

where the last "statistical term" decreases as the number of workers n grows. This provides a theoretical explanation for the benefit of distributed optimization and the use of a large number of workers.

Methods with only worker-to-server compression. However, the first "communication term" in (2) does not scale with n. It turns out that the term $d\tau$ corresponding to the w2s communication can be improved using compressed communication techniques (Seide et al., 2014; Alistarh et al., 2017). There are many efficient methods with compressed communication, including DIANA (Mishchenko et al., 2019), Accelerated DIANA (Li et al., 2020), MARINA (Gorbunov et al., 2021), and DASHA (Tyurin and Richtárik, 2023b), which rely on *unbiased compressors*.

Definition 1.5. A stochastic mapping $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is an unbiased compressor if there exists $\omega \geq 0$ such that $\mathbb{E}[\mathcal{C}(x)] = x$ and $\mathbb{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq \omega \|x\|^2$. We denote the family of such compressors by $\mathbb{U}(\omega)$. Unless otherwise stated, we assume that all compressors are statistically independent.

One of the most common examples of an unbiased compressor is Rand$K \in \mathbb{U}(d/K - 1)$, which operates by selecting K coordinates of the input vector x uniformly at random, rescaling them by d/K, and setting all other coordinates to zero (see Definition D.1). Beyond Rand$K$, a variety of other unbiased compressors have been proposed and studied in the literature (Beznosikov et al., 2020; Xu et al., 2021; Horváth et al., 2022; Szlendak et al., 2021).

Using the seminal ideas (Seide et al., 2014; Alistarh et al., 2017), we can construct the following compressed method:

$$x^{k+1} = x^k - \frac{\gamma}{nbm} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathcal{C}_{ij}^k \left(\sum_{r=1}^{b} \nabla f(x^k; \xi_{ir}^k) \right), \tag{3}$$

where worker i computes b stochastic gradients, and then sends m compressed vectors $\{\mathcal{C}_{ij}^k(\cdot)\}_{j\in[m]}$ to the server[1], which aggregates and calculates x^{k+1}. Using Rand$K$ and a proper choice of parameters, we can improve (2) to the time complexity

$$\mathcal{O}\left(\frac{d\kappa L\Delta}{\varepsilon} + \tau\left(\frac{d}{n}+1\right)\frac{L\Delta}{\varepsilon} + \sqrt{\frac{d\tau h\sigma^2}{n\varepsilon}}\frac{L\Delta}{\varepsilon} + h\left(1+\frac{\sigma^2}{n\varepsilon}\right)\frac{L\Delta}{\varepsilon} \right). \tag{4}$$

However, all the listed methods, including (3), still depend on $d\kappa L\Delta/\varepsilon$, which does not improve with n, because they send full vectors from the server to the workers. Moreover, if $\kappa = \tau$, then the compression technique does not help at all, since (4) equals $\mathcal{O}\left(\kappa dL\Delta/\varepsilon + h\left(L\Delta/\varepsilon + \sigma^2 L\Delta/n\varepsilon^2\right)\right)$, reducing to (2), as in Synchronous SGD with minibatching that does not use compression at all. The "communication term" $d\kappa L\Delta/\varepsilon$ does not decrease as the number of workers grows.

Bidirectionally compressed methods. A large body of work employs communication compression to reduce the s2w communication cost (Zheng et al., 2019; Liu et al., 2020; Philippenko and Dieuleveut, 2021; Fatkhullin et al., 2021; Yue et al., 2023; Gruntkowska et al., 2023; Tyurin and Richtárik, 2023a). However, none of these works provide theoretical guarantees in which the s2w communication term scales with n. A notable exception is the work by Gruntkowska et al. (2024), where the authors design the M3 method for the heterogeneous setting that achieves scaling under an additional assumption; however, they consider the deterministic setting, unlike our stochastic setting, which is much more challenging.

1.2 Recent pessimistic lower bound and motivation

A recent work by Tyurin (2026) analyzes our stochastic setting under Assumptions 1.1, 1.2, 1.3, and 1.4, and proves a pessimistic result that it is (informally) impossible to achieve a time complexity better than

$$\tilde{\Omega}\left(\min\left\{ \frac{d\kappa L\Delta}{\varepsilon} + h\left(1+\frac{\sigma^2}{n\varepsilon}\right)\frac{L\Delta}{\varepsilon} + \tau\left(1+\frac{d}{n}\right)\frac{L\Delta}{\varepsilon} + \sqrt{\frac{d\tau h\sigma^2}{n\varepsilon}}\frac{L\Delta}{\varepsilon}, \frac{h\sigma^2 L\Delta}{\varepsilon^2} + \frac{hL\Delta}{\varepsilon} \right\} \right) \tag{5}$$

up to logarithmic factors, using random sparsification in both the homogeneous and heterogeneous settings. In particular, if $\kappa \simeq \tau$, then the lower bound is

$$\tilde{\Omega}\left(\min\left\{ \frac{d\kappa L\Delta}{\varepsilon} + \frac{hL\Delta}{\varepsilon} + \frac{h\sigma^2 L\Delta}{n\varepsilon^2}, \frac{h\sigma^2 L\Delta}{\varepsilon^2} + \frac{hL\Delta}{\varepsilon} \right\} \right), \tag{6}$$

which can be matched by Synchronous SGD (see (2)) or by the non-distributed SGD method. This means that using methods with random sparsification for compression in the distributed stochastic centralized setting does not yield any advantage, and it is infeasible to scale both the communication and statistical terms.

> Despite the significant progress in centralized stochastic optimization, classical Synchronous SGD with minibatching or non-distributed SGD remain the state-of-the-art method in practical scenarios where the server-to-worker (s2w) communication time is non-negligible. In particular, the lower bound (6) of Tyurin (2026) proves that there is no hope for improvement under Assumptions 1.1, 1.2, 1.3, and 1.4 when $\kappa \simeq \tau$. *Is the situation truly this pessimistic, or is it possible to design more advanced methods with provably better time complexity guarantees under an additional structural assumption?*

[1] Or, almost equivalently, use one Rand$m$ compressor, since both constructions have variance parameter at most d/m.

1.3 Contributions

In this paper, we develop two new methods, Inkheart SGD and M4, that achieve new state-of-the-art time complexities in centralized stochastic optimization. Despite significant prior progress on this topic, this is the first result to improve the convergence of the naive Synchronous SGD method in the practical setting where communication from the server to the workers is not negligible and the workers have access only to stochastic gradients.

In order to achieve this result, due to (Tyurin, 2026), it is necessary to introduce an additional structural assumption about f. We consider Assumption 1.6 (Gruntkowska et al., 2024; Wang et al., 2023). *Notice that Assumption 1.6 follows from Assumption 1.1 and vice versa due to Proposition 1.7.* This assumption captures an additional structural property of f through the structural parameters L_A and L_B. It is always possible to take $L_A = L$ and $L_B = 0$. However, our theoretical improvements are most significant when L_A is small. For instance, $L_A = 0$ when f is a quadratic function, and L_A is small when f has a "slowly changing" Hessian (see Propositions 1.8 and 1.9); we also observe that L_A is small in a small-scale practical task (see Section G). We empirically verify our improvement in Section H.

New state of the art in stochastic centralized optimization. Sections 2 and 3. Under an additional assumption (necessary due to the lower bound), in both homogeneous and heterogeneous scenarios, we develop Inkheart SGD and M4 that achieve new state-of-the-art time complexities

$$\mathcal{O}\left(\max\left\{h, \tau, \kappa, \frac{d\kappa}{\sqrt{n}}, \frac{d\tau}{n}, \left(\frac{d^3\tau\kappa^2}{n}\right)^{\frac{1}{3}}, \sqrt{\frac{d\sigma^2 h\tau}{n\varepsilon}}, \frac{\sigma^2 h}{n\varepsilon}\right\}\frac{L_{\max}\Delta}{\varepsilon} + \frac{d\kappa L_A\Delta}{\varepsilon}\right) \tag{7}$$

and

$$\mathcal{O}\left(\max\left\{h, \tau, \kappa, \frac{d\kappa}{\sqrt{n}}, \frac{d\tau}{\sqrt{n}}, \left(\frac{d^3\tau^2\kappa}{n}\right)^{\frac{1}{3}}, \left(\frac{d^2\tau^2 h\sigma^2}{n\varepsilon}\right)^{\frac{1}{3}}, \frac{\sigma^2 h}{n\varepsilon}\right\}\frac{L_{\max}\Delta}{\varepsilon} + \frac{d\kappa L_A\Delta}{\varepsilon} + d\tau\right), \tag{8}$$

respectively (Theorems 2.2 and 3.5). In particular, (7) is never worse[2] than (5) due to Proposition 1.7, since we can always take $L_A = L$ and $L_B = 0$ and use that $\left({}^{d^3\tau\kappa^2}/_{n}\right)^{1/3} \leq {}^{d\tau}/_{n} + d\kappa$. *However, (7) can be arbitrarily better in the regime when n is large and L_A is small, since in this case (7) $\to \max\{h, \tau, \kappa\}{}^{L_{\max}\Delta}/_{\varepsilon}$, which does not depend on d and ${}^{\sigma^2}/_{\varepsilon}$.* We obtain similar conclusions in the heterogeneous scenario for (8), up to the initial communication cost $d\tau$, which does not depend on ε.

Extension to the setting with heterogeneous times. Section 4. We also extend our result to the scenario when the computation and communication times are heterogeneous (Assumption 4.1). This is a standard assumption in the analysis of parallel and asynchronous methods (Mishchenko et al., 2022; Tyurin and Richtárik, 2023c). In this setting, previous works have the same issue: both computation and communication terms do not scale in the complexities (Tyurin et al., 2024). In Section 4, we analyze a new method, Heterogeneous-Time Inkheart SGD, based on Inkheart SGD and prove a new state-of-the-art time complexity.

1.4 Preliminaries: Functional (L_A, L_B) Inequality

In order to improve the pessimistic lower bound, we introduce the Functional (L_A, L_B) Inequality:

Assumption 1.6 (Functional (L_A, L_B) Inequality; Gruntkowska et al. (2024); Wang et al. (2023))**.** There exist constants $L_A, L_B \geq 0$ such that

$$\left\|\frac{1}{n}\sum_{i=1}^{n}(\nabla f(x+u_i) - \nabla f(x))\right\|^2 \leq L_A^2\left(\frac{1}{n}\sum_{i=1}^{n}\|u_i\|^2\right) + L_B^2\left\|\frac{1}{n}\sum_{i=1}^{n}u_i\right\|^2 \tag{9}$$

for all $n \geq 1$ and $x, u_1, \ldots, u_n \in \mathbb{R}^d$.

[2]Here we assume that the minimum in (5) is attained by the first expression. If the second term is smaller, communication costs outweigh the benefits of parallel computation, and a non-distributed method is preferable.

The Inkheart SGD Method

Initialize vector $x^0 \in \mathbb{R}^d$ and take $x_i^0 = x^0$ for all $i \in [n]$, step size $\gamma > 0$, params $b, m, \ell \in \mathbb{N}$, compressor $\mathcal{C}_{ij}^k \in \mathbb{U}(\omega)$, and $\mathcal{C}_{\text{s},ij}^k \in \mathbb{U}(\omega_s)$ for all $i, j, k \geq 0$. Then, iterate the following steps for $k = 0, 1, \ldots$:

$$
\begin{aligned}
g^k &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{bm}\sum_{j=1}^{m}\mathcal{C}_{ij}^k\left(\sum_{r=1}^{b}\nabla f(x_i^k;\xi_{ir}^k)\right), \\
x^{k+1} &= x^k - \gamma g^k, \\
c^k &\sim \text{Bernoulli}(p), \quad \text{where } p \in (0,1] \text{ is a parameter}, \\
x_i^{k+1} &= \begin{cases} x^{k+1} & \text{if } c^k = 1, \\ x_i^k + \frac{1}{\ell}\sum_{j=1}^{\ell}\mathcal{C}_{\text{s},ij}^k(x^{k+1} - x^k) & \text{if } c^k = 0 \end{cases} \\
&\quad \text{for all } i \in [n].
\end{aligned}
\qquad \text{(Inkheart SGD)}
$$

The samples ξ_{ir}^k are mutually independent for all i, r, and $k \geq 0$, and independent of the compressors and the Bernoulli random variables.

Before we present our new methods and theoretical results, let us discuss the properties of Assumption 1.6 and its connection to Assumption 1.1. Note that Assumption 1.6 only refines the functional smoothness properties and *does not* restrict the class of considered functions; this is formalized by the following proposition.

Proposition 1.7. *If Assumption 1.1 holds, then Assumption 1.6 is satisfied with $L_A = L$ and $L_B = 0$. If Assumption 1.6 holds, then f is L-smooth with $L = \sqrt{L_A^2 + L_B^2}$.*

The next easily verified proposition says that we can choose $L_A = 0$ when f is a quadratic function.

Proposition 1.8. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a quadratic function defined as $f(x) = \frac{1}{2}x^\top \mathbf{H}x + \langle b, x\rangle + c$, where $\mathbf{H} \in \mathbb{S}^d, b \in \mathbb{R}^d$ and $c \in \mathbb{R}$. Then, Assumption 1.6 holds with $L_A = 0$ and $L_B = \|\mathbf{H}\|_2$.*

While quadratic optimization problems are popular in many applications, it is important to understand when L_A is small for non-quadratic functions. We also consider the proposition below.

Proposition 1.9 (Gruntkowska et al. (2024))**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable, L–smooth, and $D \in \mathbb{R}$ be the smallest constant such that $\sup_{z_1,\ldots,z_n \in \mathbb{R}^d}\left\|\nabla^2 f(z_1) - \frac{1}{n}\sum_{j=1}^{n}\nabla^2 f(z_j)\right\| \leq D$ for all $n \geq 1$. Then, Assumption 1.6 holds with $L_A = \sqrt{2}D$ and $L_B = \sqrt{2}L$. Moreover, D is always finite, and $D \leq 2L$.*

This proposition states that if the Hessian does not change significantly, then L_A is small. In the worst case, $L_A = \sqrt{2}D \leq 2\sqrt{2}L$; however, it can be arbitrarily smaller than L, for instance, $L_A = 0$ when f is a quadratic function, or consider the following example of an additive non-quadratic function with a small smoothness constant.

Example 1.10. Let $f : \mathbb{R}^d \to \mathbb{R}$ such that $f(x) = \frac{1}{2}x^\top \mathbf{H}x + \langle b, x\rangle + c + g(x)$, where $\mathbf{H} \in \mathbb{S}^d, b \in \mathbb{R}^d, c \in \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ is L_{small}–smooth. Then, f satisfies Assumption 1.6 with $L_A = 2\sqrt{2}L_{\text{small}}$ and $L_B = \sqrt{2}(\|\mathbf{H}\| + L_{\text{small}})$.

Numerical experiments. In Section G, we also consider a machine learning task with a convolutional neural network, where we provide empirical evidence that L_A is relatively small and can be $20 \sim 200$ times smaller than L.

2 Inkheart SGD: A New Algorithm in a Homogeneous Setting

We now present our new method, Inkheart SGD, which combines ideas from the classical QSGD method (Alistarh et al., 2017) and a recent paper (Gruntkowska et al., 2024). Let us explain how

Inkheart SGD works. At the beginning of every iteration k, each worker $i \in [n]$ computes a mini-batch of size b. These estimates are summed locally and then compressed using the operators $\{\mathcal{C}_{ij}^k\}_{j\in[m]}$. Then, they are transmitted to the server, which averages them across all workers to form g^k. Next, the server performs a standard gradient descent step $x^{k+1} = x^k - \gamma g^k$. After that, the algorithm decides whether to synchronize all local models with the updated global iterate or to perform a compressed update. In particular, with probability $p \in (0, 1]$, all workers set $x_i^{k+1} = x^{k+1}$, corresponding to a full synchronization step. Otherwise, with probability $1 - p$, each worker updates its local model using compressed vectors $\mathcal{C}_{\mathrm{s},ij}^k(x^{k+1} - x^k)$. In this case, worker i adds the average of these ℓ compressed vectors to the local x_i^k. Notice that the value of p is small, non-compressed vectors are sent rarely, and this does not have an adverse effect on the final time complexity.

2.1 Theoretical results

For Inkheart SGD, we can prove the following *iteration* rate:

Theorem 2.1 (Iteration complexity. Follows from Theorem E.1)**.** *Let Assumptions 1.1, 1.2, 1.3, and 1.6 be satisfied. Then, Inkheart SGD with* $\gamma = \Theta\left(1/\max\left\{L_{\max}, \frac{\sqrt{\omega\omega_s}L_{\max}}{\sqrt{pnm\ell}}, \frac{\sqrt{\omega_s}L_{\max}}{\sqrt{p\ell n}}, \frac{\sqrt{\omega_s}L_A}{\sqrt{p\ell}}\right\}\right)$, *and all* m *and* b *such that* $8\omega/m \leq n$ *and* $8\omega\sigma^2/mb\varepsilon + 8\sigma^2/b\varepsilon \leq n$ *finds an* ε*–stationary point of* (1) *after at most*

$$K = \Theta\left(\max\left\{1, \frac{\sqrt{\omega\omega_s}}{\sqrt{pnm\ell}}, \frac{\sqrt{\omega_s}}{\sqrt{p\ell n}}\right\}\frac{L_{\max}\Delta}{\varepsilon} + \frac{\sqrt{\omega_s}L_A\Delta}{\sqrt{p\ell}\varepsilon}\right) \tag{10}$$

iterations, where $L_{\max} := \max\{L, L_A, L_B\}$.

Note that Theorem 2.1 is an *auxiliary* result and does not yield a time complexity that allows us to determine whether the method can improve the pessimistic lower bounds (5) and (6). Theorem 2.1 works with all unbiased compressors that satisfy Definition 1.5. Without loss of generality, and for simplicity, we fix the Rand1 compressor with $\omega = \omega_s = d - 1$. Since this compressor sends only one coordinate, the time required to send one compressed vector from the workers to the server and vice versa is τ and κ, respectively, according to Assumption 1.4. There are three main bottlenecks in (Inkheart SGD): stochastic gradient computations and communication between the workers and the server (in both directions). The idea is to assign a time budget t to each operation; thus, $b = \left\lfloor \frac{t}{h} \right\rfloor$, $m = \left\lfloor \frac{t}{\tau} \right\rfloor$, and $\ell = \left\lfloor \frac{t}{\kappa} \right\rfloor$. Substituting ω and ω_s into (10), and noting that each iteration takes at most $\bar{t} := hb + \tau m + (pd + (1-p)\ell)\kappa$ seconds on average, it remains to minimize the total time $K \times \bar{t}$ over t to obtain the following result.

Theorem 2.2 (Time Complexity of Inkheart SGD)**.** *Consider the assumptions and result of Theorem 2.1. Additionally, assume that Assumption 1.4 is satisfied and the workers and the server use Rand*1 *compressor (Definition 1.5). Let* $b = \left\lfloor \frac{t}{h} \right\rfloor$, $m = \left\lfloor \frac{t}{\tau} \right\rfloor$, *and* $\ell = \left\lfloor \frac{t}{\kappa} \right\rfloor$, *where* $t = \max\left\{h, \tau, \kappa, 16\omega\tau/n, 16\sigma^2 h/n\varepsilon, 2d\kappa/\sqrt{n}, \sqrt{32d\sigma^2 h\tau/n\varepsilon}, \left(8d^3\tau\kappa^2/n\right)^{\frac{1}{3}}\right\}$. *Then, the expected time for Inkheart SGD with* $p = \min\left\{\frac{\ell}{d}, 1\right\}$ *to find an* ε*–stationary point of* (1) *is*

$$\mathcal{O}\left(\max\left\{h, \tau, \kappa, \frac{d\kappa}{\sqrt{n}}, \frac{d\tau}{n}, \left(\frac{d^3\tau\kappa^2}{n}\right)^{\frac{1}{3}}, \sqrt{\frac{d\sigma^2 h\tau}{n\varepsilon}}, \frac{\sigma^2 h}{n\varepsilon}\right\}\frac{L_{\max}\Delta}{\varepsilon} + \frac{d\kappa L_A\Delta}{\varepsilon}\right). \tag{11}$$

Notice that, except for the last term $d\kappa L_A\Delta/\varepsilon$, all other terms with d and σ^2/ε decrease as the number of workers increases. In particular, in the case when $L_A \approx 0$, for instance, when f is a quadratic function or L_A is small due to Proposition 1.9, this complexity tends to $\mathcal{O}\left(\max\{h, \tau, \kappa\} L_{\max}\Delta/\varepsilon\right)$, which does not depend on d or σ^2/ε. To the best of our knowledge, this is the first result in distributed stochastic optimization. When $\kappa = \tau$, the time complexity equals $\mathcal{O}\left(\max\left\{h, \kappa, \sqrt{d\sigma^2 h\kappa/n\varepsilon}, d\kappa/n^{1/3}, \sigma^2 h/n\varepsilon\right\} L_{\max}\Delta/\varepsilon + d\kappa L_A\Delta/\varepsilon\right)$. Due to Proposition 1.7, it is never worse than $\mathcal{O}\left(d\kappa L\Delta/\varepsilon + hL\Delta/\varepsilon + h\sigma^2 L\Delta/n\varepsilon^2\right)$ achieved by Synchronous SGD with mini-batching. However, the former can be arbitrarily smaller when n is large and L_A is small. Theorem 2.2 is proved for Rand1. Similarly, one can extend it to any other unbiased compressor, where the only

The M4 Method
(M4 = MARINA-P + Momentum + Momentum + MARINA):

Initialize vector $x^0 \in \mathbb{R}^d$ and vectors $v_i^0 \in \mathbb{R}^d$ for all $i \in [n]$ and take $g^0 = {}^1\!/_n \sum_{i=1}^n v_i^0$ and $w_i^0 = x_i^0 = x^0$ for all $i \in [n]$, step size $\gamma > 0$, probabilities $0 < p_{\rm s}, p \leq 1$ and params $b \in \mathbb{N}$, $\nu, \mu \in (0, 1]$, compressor $\mathcal{C}_i^k \in \mathbb{U}(\omega)$, and $\mathcal{C}_{{\rm s},i}^k \in \mathbb{U}(\omega_s)$ for all $i, k \geq 0$. Then, iterate the following steps for $k = 0, 1, \ldots$:

$$
\begin{aligned}
x^{k+1} &= x^k - \gamma g^k, \\
w_i^{k+1} &= \begin{cases} x^{k+1} & \text{with probability } p_{\rm s}, \\ w_i^k + \mathcal{C}_{{\rm s},i}^k(x^{k+1} - x^k) & \text{with probability } 1 - p_{\rm s}, \end{cases} \qquad \forall i \in [n] \\
x_i^{k+1} &= (1-\mu)x_i^k + \mu w_i^{k+1} \qquad \forall i \in [n] \\
v_i^{k+1} &= (1-\nu)v_i^k + \frac{\nu}{b}\sum_{r=1}^{b} \nabla f_i(x_i^{k+1}; \xi_{ir}^{k+1}) \qquad \forall i \in [n] \\
g^{k+1} &= \begin{cases} \frac{1}{n}\sum_{i=1}^n v_i^{k+1} & \text{with probability } p, \\ g^k + \frac{1}{n}\sum_{i=1}^n \mathcal{C}_i^k(v_i^{k+1} - v_i^k) & \text{with probability } 1 - p \end{cases}
\end{aligned}
\tag{M4}
$$

where the first probabilistic choice is the same for all workers from set $[n]$: one Bernoulli random variable is drawn for all workers. The coins for the first and second probabilistic choices with $p_{\rm s}$ and p are independent. The samples ξ_{ir}^k are mutually independent for all i, r, and $k \geq 0$, and independent of the compressors and the Bernoulli random variables.

things that would change are the values of ω and ω_s and the amount required to transmit one compressed value.

3 M4: A New Algorithm in a Heterogeneous Setting

In the heterogeneous setting, we consider the following nonconvex distributed optimization task:

$$
\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{n}\sum_{i=1}^n f_i(x) \right\}. \tag{12}
$$

Instead of Assumptions 1.1, 1.3, and 1.6, we consider the following assumptions in the heterogeneous setting.

Assumption 3.1. The function f is L–smooth, and the function f_i is L_i smooth for all $i \in [n]$. We define $\hat{L}^2 := \frac{1}{n}\sum_{i=1}^n L_i^2$.

Assumption 3.2 (Heterogeneous setting)**.** For all $i \in [n]$ and $x \in \mathbb{R}^d$, the stochastic gradients $\nabla f_i(x;\xi)$ are unbiased and have σ^2–bounded variance: $\mathbb{E}_\xi\left[\nabla f_i(x;\xi)\right] = \nabla f_i(x)$ and $\mathbb{E}_\xi[\|\nabla f_i(x;\xi) - \nabla f_i(x)\|^2] \leq \sigma^2$ for all $x \in \mathbb{R}^d$, where $\sigma > 0$ is some constant.

Assumption 3.3 (Functional (L_A, L_B) similarity)**.** There exist constants $L_A, L_B \geq 0$ such that

$$
\left\| \frac{1}{n}\sum_{i=1}^n (\nabla f_i(x + u_i) - \nabla f_i(x)) \right\|^2 \leq L_A^2 \left(\frac{1}{n}\sum_{i=1}^n \|u_i\|^2 \right) + L_B^2 \left\| \frac{1}{n}\sum_{i=1}^n u_i \right\|^2 \tag{13}
$$

for all $n \geq 1$ and $x, u_1, \ldots, u_n \in \mathbb{R}^d$. We define $L_{\max} := \max\left\{\max_{i\in[n]} L_i, L_A, L_B\right\}$.

We are ready to present our new method, M4. The design of this method is based on the ideas from (Gorbunov et al., 2021; Fatkhullin et al., 2023; Gruntkowska et al., 2024). At each iteration k, the server first performs a gradient descent step $x^{k+1} = x^k - \gamma g^k$. Then, with probability $p_{\rm s}$, all workers synchronize by setting $w_i^{k+1} = x^{k+1}$, while with probability $1 - p_{\rm s}$, they update w_i^{k+1} using a compressed version of the model difference $x^{k+1} - x^k$ with compressor $\mathcal{C}_{{\rm s},i}^k$. The workers receive either x^{k+1} or the compressed vectors from the server. Next, each worker forms its local

iterate x_i^{k+1} as a convex combination of x_i^k and w_i^{k+1} with weight μ, and updates its local gradient estimator v_i^{k+1} using a mini-batch of size b and momentum parameter ν. Finally, the server updates the global estimator g^{k+1}: with probability p, it computes the exact average $\frac{1}{n}\sum_{i=1}^n v_i^{k+1}$, receiving the full vectors from the workers, and with probability $1-p$, it performs a compressed update based on compressor $\mathcal{C}_i^k$ applied to the differences $v_i^{k+1}-v_i^k$ that the workers send to the server. Similarly to Inkheart SGD, since p and $p_{\rm s}$ are small, the synchronization of full vectors does not slow down the optimization procedure on average. Without loss of generality, we assume that the workers and the server send one compressed vector, since one can easily show that the average of unbiased compressors is an unbiased compressor.

3.1 Theoretical results

Theorem 3.4 (Iteration complexity). *Let Assumptions 3.1, 1.2, 3.2, and 3.3 be satisfied. Then, M4 with momentum parameters* $\nu=\mu=\eta:=\min\left\{\frac{1}{6}\sqrt{\frac{bn\varepsilon}{\omega(\omega+1)\sigma^2}},\frac{bn\varepsilon}{12\sigma^2},\left(\frac{n}{\omega(\omega+1)\omega_{\rm s}}\right)^{\frac{1}{3}},1\right\}$, $p_{\rm s}=\frac{1}{\omega_{\rm s}+1}, p=\frac{1}{\omega+1}$, $v_i^0=\frac{1}{b_{\rm init}}\sum_{b=1}^{b_{\rm init}}\nabla f_i(x_i^0;\xi_{i,b}^0)$ *where* $b_{\rm init}=\left\lceil\frac{b}{\nu}\right\rceil$, *and step size* $\gamma=\Theta\left(\sqrt{\left(\omega_{\rm s}(\omega_{\rm s}+1)L_A^2+\frac{\omega_{\rm s}}{n}(\omega_{\rm s}+1)L_B^2+\left(\frac{\omega(\omega+1)}{n}+\frac{1}{\eta^2}\right)L_{\max}^2\right)}\right)^{-1}$ *finds an* ε*–stationary point of* (12) *after at most*

$$K=\mathcal{O}\left(\frac{\Delta}{\gamma\varepsilon}+\max\left\{\sqrt{\frac{\omega(\omega+1)\sigma^2}{bn\varepsilon}},\frac{\sigma^2}{bn\varepsilon},\left(\frac{\omega(\omega+1)\omega_{\rm s}}{n}\right)^{\frac{1}{3}},1\right\}\right)$$

iterations.

In the parameter choices, a constraint with zero denominator is interpreted as $+\infty$. In the proof, we define the Lyapunov function Ψ^k as follows:

$$\Psi^k=f(x^k)-f^*+\lambda_A\left\|g^k-v^k\right\|^2+\lambda_B\left\|v^k-\nabla f(x^k)\right\|^2+\frac{\lambda_C}{n}\sum_{i=1}^n\left\|v_i^k-\nabla f_i(x_i^k)\right\|^2$$
$$+\frac{\lambda_D}{n}\sum_{i=1}^n\left\|w_i^k-x_i^k\right\|^2+\frac{\lambda_E}{n}\sum_{i=1}^n\left\|w_i^k-x^k\right\|^2+\lambda_F\left\|\frac{1}{n}\sum_{i=1}^n\left(w_i^k-x_i^k\right)\right\|^2+\lambda_G\left\|\frac{1}{n}\sum_{i=1}^n w_i^k-x^k\right\|^2,$$

where $\lambda_A,\lambda_B,\lambda_C,\lambda_D,\lambda_E,\lambda_F,\lambda_G$ are defined in Theorem F.9. Using this choice of coefficients, we prove the inequality $\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla f(x^k)\|^2\right]\leq 2\mathbb{E}[\Psi^0]/\gamma K+\frac{\varepsilon}{2}$, and then, choosing $w_i^0=x_i^0=x^0$ and $g^0=1/n\sum_{i=1}^n v_i^0$, we obtain $\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla f(x^k)\|^2\right]\leq 2\Delta/\gamma K+\varepsilon/2+2/\gamma K\left(\lambda_B\mathbb{E}\left[\left\|v^0-\nabla f(x^0)\right\|^2\right]+\lambda_C\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|v_i^0-\nabla f_i(x_i^0)\right\|^2\right]\right)$. In order to reduce the dependence on the last two terms, we cannot set $v_i^0=\nabla f_i(x_i^0)$ in the stochastic setting; instead, we initialize it with a minibatch of size $b_{\rm init}$.

As in the homogeneous case, Theorem 3.4 is an auxiliary result, and we now present our main theoretical result in the heterogeneous setting.

Theorem 3.5 (Time Complexity). *Consider the assumptions and result of Theorem 3.4. Additionally, assume that Assumption 1.4 is satisfied and the workers and server use RandK compressor with* $K=\min\{d,\lfloor\frac{t}{\tau}\rfloor\}$ *and* $K=\min\{d,\lfloor\frac{t}{\kappa}\rfloor\}$, *respectively, and use batch size* $b=\lfloor\frac{t}{h}\rfloor$, *where* $t=\max\left\{h,\tau,\kappa,\sqrt[3]{d^2\tau^2h\sigma^2/n\varepsilon}\right\}$. *Then, the expected time for M4 to find an* ε*–stationary point of* (12) *is at most*

$$\mathcal{O}\left(\max\left\{h,\tau,\kappa,\frac{d\kappa}{\sqrt{n}},\frac{d\tau}{\sqrt{n}},\left(\frac{d^3\tau^2\kappa}{n}\right)^{\frac{1}{3}},\left(\frac{d^2\tau^2h\sigma^2}{n\varepsilon}\right)^{\frac{1}{3}},\frac{\sigma^2h}{n\varepsilon}\right\}\frac{L_{\max}\Delta}{\varepsilon}+\frac{d\kappa L_A\Delta}{\varepsilon}+d\tau\right)\tag{14}$$

seconds, including the initial computation of minibatches of size $b_{\rm init}$ *and the initial transmission of the vectors* v_i^0 *to the server.*

The dependence here is slightly worse than in (11), the homogeneous setting. Nevertheless, up to the initial communication cost $d\tau$, which does not depend on ε, the conclusion is the same, and M4 enjoys scaling with n when the structural parameter L_A is small, unlike all previous compressed and non-compressed methods in stochastic optimization.

4 Extension to Optimization with Heterogeneous Times

In this section, we extend the result from Section 2 to the case where computation and communication times are heterogeneous, an assumption used to compare asynchronous methods (Mishchenko et al., 2022; Tyurin and Richtárik, 2023c).

Assumption 4.1. For all $i \in [n]$, worker i requires at most h_i seconds to compute a stochastic gradient, and communication from the server to any worker i takes at most κ_i seconds per coordinate, and communication from worker i to the server takes at most τ_i seconds per coordinate.

In Section B, we extend Inkheart SGD to this setting. There are two essential changes: i) the number of computed stochastic gradients and sent compressed vectors varies across workers; ii) the received compressed vectors are aggregated with a particular optimal choice of weights (17) to minimize the variance of g^k.

4.1 Theoretical results

Theorem 4.2 (Time Complexity of Heterogeneous-Time Inkheart SGD)**.** *Let Assumptions 1.1, 1.2, 1.3, and 1.6 be satisfied. Additionally, assume that Assumption 4.1 is satisfied and the workers and the server use Rand1 compressors. Let* $b_i = \left\lfloor \frac{t}{h_i} \right\rfloor, m_i = \left\lfloor \frac{t}{\tau_i} \right\rfloor, \ell_i = \left\lfloor \frac{t}{\kappa_i} \right\rfloor$, *where* $\ell_{\min} = \left\lfloor \frac{t}{\kappa_{\max}} \right\rfloor$, $t = \max\left\{\max_{i\in[n]}\left\{h_i, \tau_i, \kappa_i\right\}, s^*\right\}$, $\kappa_{\max} = \max_{i\in[n]} \kappa_i$, *and* s^* *is the solution of the equation*

$$\left(\sum_{i=1}^{n} \frac{1}{\frac{16\omega\tau_i}{s} + \frac{16\sigma^2 h_i}{\varepsilon s} + \frac{32\sigma^2\omega h_i\tau_i}{\varepsilon s^2} + \frac{4d\omega_{\mathrm{s}}\kappa_{\max}\kappa_i}{s^2} + \frac{8d\omega_{\mathrm{s}}\omega\kappa_{\max}\kappa_i\tau_i}{s^3}}\right)^{-1} = 1. \tag{15}$$

Then, the expected time for (Heterogeneous-Time Inkheart SGD) *with* $p = \min\left\{\frac{\ell_{\min}}{d}, 1\right\}$ *to find an* ε*–stationary point of* (1) *is at most*

$$\mathcal{O}\left(\frac{tL_{\max}\Delta}{\varepsilon} + \frac{d\kappa_{\max}L_A\Delta}{\varepsilon}\right). \tag{16}$$

When the times are equal, this result reduces to Theorem 2.2, and we can obtain an explicit formula that exhibits scaling with n when L_A is small. In general, following (Tyurin et al., 2024), to find the parameters and derive the final time complexity, one has to solve (15). When the times κ_i, h_i, τ_i do not grow too fast, one can show that $s^* \to 0$ as $n \to \infty$, and the limiting time complexity does not depend on d, unlike the result by (Tyurin et al., 2024).

Notice that both t and s^* depend on $\max_{i\in[n]}\left\{h_i, \tau_i, \kappa_i\right\}$ and $\kappa_{\max}$. This might be a problem when, for instance, the computation and communication time of one of the workers is huge. For instance, if $\kappa_n \to \infty$, then $\kappa_{\max} \to \infty$ and (16) $\to \infty$.

A more robust strategy is to take a subset of workers $\emptyset \neq S \subseteq [n]$, evaluate the time $t(S) := \max\left\{\max_{i\in S}\left\{h_i, \tau_i, \kappa_i\right\}, s^*(S)\right\}$, where $s^*(S)$ is the solution of (15) when only the workers in S participate in the optimization ($\kappa_{\max} \to \max_{i\in S}\kappa_i$ and $\sum_{i=1}^{n} \to \sum_{i\in S}$), and then minimize the time complexity (16) over all subsets $\emptyset \neq S \subseteq [n]$. This way, we find the fastest subset of workers S^* and obtain a better time complexity

$$\mathcal{O}\left(\frac{t(S^*)L_{\max}\Delta}{\varepsilon} + \frac{d\max\limits_{i\in S^*}\kappa_i L_A\Delta}{\varepsilon}\right).$$

In Section E.2, we describe an efficient polynomial-time algorithm that minimizes the complexity over all subsets $\emptyset \neq S \subseteq [n]$.

Theorem 4.3. *A subset of workers S^* that minimizes*

$$\tilde{T}(S) := \max\left\{t(S)L_{\max}, d\max_{i\in S}\kappa_i L_A\right\}$$

over all $\emptyset \neq S \subseteq [n]$ can be found using Algorithm 1.

5 Conclusion and Future Work

The centralized stochastic optimization with compressed methods is a well-explored direction (e.g., (Wu et al., 2018; Mishchenko et al., 2019; Gorbunov et al., 2021; Fatkhullin et al., 2021; Tyurin and Richtárik, 2023b; Huang et al., 2022; Tyurin and Richtárik, 2023a)). Nevertheless, for the first time, we develop two new algorithms, Inkheart SGD and M4, that provably improve upon simple baselines, non-distributed SGD and Synchronous SGD, in the stochastic centralized setting. A similar improvement is achieved in the presence of heterogeneous computation and communication times. To present this improvement, we rely on structural assumptions that are necessary due to the lower bound of (Tyurin, 2026) and do not restrict the class of functions. We show that this assumption is both theoretically sound and practically relevant. One interesting direction is to consider alternative assumptions that capture the structure of f and could potentially improve the baselines in stochastic distributed optimization.

Acknowledgements

The work was supported by the grant for research centers in the field of AI provided by the Ministry of Economic Development of the Russian Federation in accordance with the agreement 000000C313925P4F0002 and the agreement №139-10-2025-033.

References

Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. (2017). QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1709–1720.

Beznosikov, A., Horváth, S., Richtárik, P., and Safaryan, M. (2020). On biased compression for distributed learning. *arXiv preprint arXiv:2002.12410*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Fatkhullin, I., Sokolov, I., Gorbunov, E., Li, Z., and Richtárik, P. (2021). EF21 with bells & whistles: Practical algorithmic extensions of modern error feedback. *arXiv preprint arXiv:2110.03294*.

Fatkhullin, I., Tyurin, A., and Richtárik, P. (2023). Momentum provably improves error feedback! *Advances in Neural Information Processing Systems*.

Gorbunov, E., Burlachenko, K. P., Li, Z., and Richtárik, P. (2021). Marina: Faster non-convex distributed learning with compression. In *International conference on machine learning*, pages 3788–3798. PMLR.

Gruntkowska, K., Tyurin, A., and Richtárik, P. (2023). EF21-P and friends: Improved theoretical communication complexity for distributed optimization with bidirectional compression. In *International Conference on Machine Learning*, pages 11761–11807. PMLR.

Gruntkowska, K., Tyurin, A., and Richtárik, P. (2024). Improving the worst-case bidirectional communication complexity for nonconvex distributed optimization under function similarity. *Advances in Neural Information Processing Systems*, 37:88807–88873.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Horváth, S., Ho, C.-Y., Horváth, v., Sahu, A. N., Canini, M., and Richtárik, P. (2022). Natural compression for distributed deep learning. In *Mathematical and Scientific Machine Learning*, pages 129–141. PMLR.

Huang, J., Qian, F., Gerber, A., Mao, Z. M., Sen, S., and Spatscheck, O. (2012). 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.

Huang, X., Chen, Y., Yin, W., and Yuan, K. (2022). Lower bounds and nearly optimal algorithms in distributed learning with communication compression. *Advances in Neural Information Processing Systems*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

Krizhevsky, A., Nair, V., and Hinton, G. (2010). Cifar-10 (canadian institute for advanced research). *URL http://www. cs. toronto. edu/kriz/cifar. html*, 5(4):1.

Lan, G. (2020). *First-order and stochastic optimization methods for machine learning*. Springer.

LeCun, Y., Cortes, C., and Burges, C. (2010). Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2.

Li, Z., Kovalev, D., Qian, X., and Richtárik, P. (2020). Acceleration for compressed gradient descent in distributed and federated optimization. In *International Conference on Machine Learning*.

Liu, X., Li, Y., Tang, J., and Yan, M. (2020). A double residual compression algorithm for efficient distributed learning. In *International Conference on Artificial Intelligence and Statistics*, pages 133–143. PMLR.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Mishchenko, K., Bach, F., Even, M., and Woodworth, B. (2022). Asynchronous SGD beats minibatch SGD under arbitrary delays. *Advances in Neural Information Processing Systems*.

Mishchenko, K., Gorbunov, E., Takáč, M., and Richtárik, P. (2019). Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*.

Narayanan, A., Zhang, X., Zhu, R., Hassan, A., Jin, S., Zhu, X., Zhang, X., Rybkin, D., Yang, Z., Mao, Z. M., et al. (2021). 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.

Philippenko, C. and Dieuleveut, A. (2021). Preserved central model for faster bidirectional compression in distributed settings. *Advances in Neural Information Processing Systems*, 34:2387–2399.

Seide, F., Fu, H., Droppo, J., Li, G., and Yu, D. (2014). 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*.

Szlendak, R., Tyurin, A., and Richtárik, P. (2021). Permutation compressors for provably faster distributed nonconvex optimization. In *International Conference on Learning Representations*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tyurin, A. (2026). Proving the limited scalability of centralized distributed optimization via a new lower bound construction. In *International Conference on Learning Representations (ICLR)*.

Tyurin, A., Pozzi, M., Ilin, I., and Richtárik, P. (2024). Shadowheart SGD: Distributed asynchronous SGD with optimal time complexity under arbitrary computation and communication heterogeneity. *Advances in Neural Information Processing Systems*, 37.

Tyurin, A. and Richtárik, P. (2023a). 2Direction: Theoretically faster distributed training with bidirectional communication compression. *Advances in Neural Information Processing Systems*.

Tyurin, A. and Richtárik, P. (2023b). DASHA: Distributed nonconvex optimization with communication compression, optimal oracle complexity, and no client synchronization. *11th International Conference on Learning Representations (ICLR)*.

Tyurin, A. and Richtárik, P. (2023c). Optimal time complexities of parallel stochastic optimization methods under a fixed computation model. *Advances in Neural Information Processing Systems*.

Wang, J., Wang, S., Chen, R.-R., and Ji, M. (2023). A new theoretical perspective on data heterogeneity in federated optimization. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*.

Wu, J., Huang, W., Huang, J., and Zhang, T. (2018). Error compensated quantized SGD and its applications to large-scale distributed optimization. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5325–5333, Stockholmsmässan, Stockholm Sweden. PMLR.

Xu, H., Ho, C.-Y., Abdelmoniem, A. M., Dutta, A., Bergou, E. H., Karatsenidis, K., Canini, M., and Kalnis, P. (2021). Grace: A compressed communication framework for distributed machine learning. In *2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS)*, pages 561–572. IEEE.

Yue, P., Zhao, H., Fang, C., He, D., Wang, L., Lin, Z., and Zhu, S.-c. (2023). Core: Common random reconstruction for distributed optimization with provable low communication complexity. *arXiv preprint arXiv:2309.13307*.

Zheng, S., Huang, Z., and Kwok, J. (2019). Communication-efficient distributed blockwise momentum SGD with error-feedback. *Advances in Neural Information Processing Systems*, 32.

Contents

1 Introduction **1**

1.1 Related work 2

1.2 Recent pessimistic lower bound and motivation 3

1.3 Contributions 4

1.4 Preliminaries: Functional (L_A, L_B) Inequality 4

2 Inkheart SGD: A New Algorithm in a Homogeneous Setting **5**

2.1 Theoretical results 6

3 M4: A New Algorithm in a Heterogeneous Setting **7**

3.1 Theoretical results 8

4 Extension to Optimization with Heterogeneous Times **9**

4.1 Theoretical results 9

5 Conclusion and Future Work **10**

A Notations **15**

B Inkheart SGD Method with Heterogeneous Computations and Communications **15**

C Properties of Functional Inequality and Proofs **16**

D Auxiliary Lemmas and Definitions **17**

E Proofs for Heterogeneous-Time Inkheart SGD and Inkheart SGD **19**

E.1 Time Complexity 22

E.2 Optimal subset of the workers 25

E.3 Bounding Variances: Auxiliary Lemmas 27

F Proofs for M4 **30**

F.1 Known lemmas 31

F.2 Auxiliary lemmas 31

F.3 Bounding variances 33

F.4 Time complexity 48

G Numerical Estimation of L_A from Assumption 1.6 **50**

H Experiments **52**

H.1 Homogeneous Block-Regularized Quadratic 52

H.2 Heterogeneous Quadratic Task 55

H.3 Homogeneous Small-Scale Machine Learning Task 58

H.4 Heterogeneous Small-Scale Machine Learning Task 58

H.5 Homogeneous CIFAR-10 experiments with ResNet-18 60

A Notations

Symbol	Description
$[n]$	Set $\{1, \ldots, n\}$
$\mathbb{R}^d$	d-dimensional Euclidean space
$\mathbb{S}^d$	Space of symmetric $d \times d$ matrices
$\|x\|$	Euclidean norm (vectors)
$\|\mathbf{H}\|$	Spectral norm (matrices)
n	Number of workers
$\mathcal{C}_i^k$	Worker-to-server compressor (worker i, iteration k)
$\mathcal{C}_{\mathrm{s},i}^k$	Server-to-worker compressor (worker i, iteration k)
$g = \mathcal{O}(f)$	There exists $C > 0$ such that $g(z) \leq C\, f(z)$ for all $z \in \mathcal{Z}$.
$g = \Omega(f)$	There exists $C > 0$ such that $g(z) \geq C\, f(z)$ for all $z \in \mathcal{Z}$.
$g = \Theta(f)$	There exist $C_1, C_2 > 0$ such that $C_1 f(z) \leq g(z) \leq C_2 f(z)$ for all $z \in \mathcal{Z}$.
$\tilde{\mathcal{O}}$, $\tilde{\Omega}$, and $\tilde{\Theta}$	The same as $\mathcal{O}$, Ω, and Θ, but up to logarithmic factors.
Δ	Initial optimality gap, $\Delta := f(x^0) - f^*$.
h	Time to compute one stochastic gradient.
τ	Time to transmit one coordinate from a worker to the server.
κ	Time to transmit one coordinate from the server to a worker.
h_i, τ_i, κ_i	Corresponding times for worker i in the heterogeneous-time setting.

B Inkheart SGD Method with Heterogeneous Computations and Communications

In this section, we provide an algorithm that extends Inkheart SGD from Section 2. See the description in Section 4.

The Heterogeneous-Time Inkheart SGD Method
(with Heterogeneous Computations and Communications)

Initialize vector $x^0 \in \mathbb{R}^d$ and take $x_i^0 = x^0$ for all $i \in [n]$, step size $\gamma > 0$, params $\{b_i, m_i, \ell_i\}_{i\in[n]}$, compressor $\mathcal{C}_{ij}^k \in \mathbb{U}(\omega)$, and $\mathcal{C}_{\mathrm{s},ij}^k \in \mathbb{U}(\omega_s)$ for all $i, j, k \geq 0$. Choose the weights

$$\beta_i = \frac{w_i}{\sum\limits_{j=1}^n w_j}, \quad \text{where} \quad w_i = \left(\frac{8\omega}{m_i} + \frac{8\sigma^2\omega}{\varepsilon b_i m_i} + \frac{8\sigma^2}{\varepsilon b_i} + \frac{\omega_s\omega}{p m_i \ell_i} + \frac{\omega_s}{p\ell_i}\right)^{-1} \quad \text{for all } i \in [n]. \tag{17}$$

Then, iterate the following steps for $k = 0, 1, \ldots$:

$$\begin{aligned}
g^k &= \sum_{i=1}^n \frac{\beta_i}{b_i m_i} \sum_{j=1}^{m_i} \mathcal{C}_{ij}^k \left(\sum_{r=1}^{b_i} \nabla f(x_i^k; \xi_{ir}^k)\right), \\
x^{k+1} &= x^k - \gamma g^k, \\
c^k &\sim \text{Bernoulli}(p), \text{ where } p \in (0, 1] \text{ is a parameter}, \\
x_i^{k+1} &= \begin{cases} x^{k+1} & \text{if } c^k = 1, \\ x_i^k + \frac{1}{\ell_i} \sum\limits_{j=1}^{\ell_i} \mathcal{C}_{\mathrm{s},ij}^k (x^{k+1} - x^k) & \text{if } c^k = 0. \end{cases}
\end{aligned}$$

(Heterogeneous-Time Inkheart SGD)

The samples ξ_{ir}^k are mutually independent for all i, r, and $k \geq 0$, and independent of the compressors and the Bernoulli random variables.

C Properties of Functional Inequality and Proofs

Proposition C.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a quadratic function defined as $f(x) = \frac{1}{2}x^\top \mathbf{H}x + \langle b, x\rangle + c$, where $\mathbf{H} \in \mathbb{S}^d, b \in \mathbb{R}^d$ and $c \in \mathbb{R}$. Then, Assumption 1.6 holds with $L_A = 0$ and $L_B = \|\mathbf{H}\|_2$.*

Proof. Note that $\nabla f(x) = \mathbf{H}x + b$ and

$$\left\|\frac{1}{n}\sum_{i=1}^n(\nabla f(x+u_i) - \nabla f(x))\right\|^2$$
$$= \left\|\frac{1}{n}\sum_{i=1}^n(\mathbf{H}(x+u_i) + b - Hx - b)\right\|^2 = \left\|\frac{1}{n}\sum_{i=1}^n \mathbf{H}u_i\right\|^2 \leq \|\mathbf{H}\|^2 \left\|\frac{1}{n}\sum_{i=1}^n u_i\right\|^2.$$

□

In the proofs, we use the following auxiliary assumption which is equivalent to Assumption 1.6 due to Theorem C.3.

Assumption C.2 (Weighted Functional (L_A, L_B) Inequality)**.** There exist constants $L_A, L_B \geq 0$ such that

$$\left\|\sum_{i=1}^n \beta_i(\nabla f(x+u_i) - \nabla f(x))\right\|^2 \leq L_A^2\left(\sum_{i=1}^n \beta_i \|u_i\|^2\right) + L_B^2\left\|\sum_{i=1}^n \beta_i u_i\right\|^2 \tag{18}$$

for all $n \geq 1$, $x, u_1, \ldots, u_n \in \mathbb{R}^d$, and $\beta_1, \ldots, \beta_n \in \mathbb{R}$ such that $0 \leq \beta_i \leq 1$ for all $i \in [n]$ and $\sum_{i=1}^n \beta_i = 1$.

Theorem C.3. *The function f satisfies Assumption 1.6 if and only if f satisfies Assumption C.2.*

Proof. Assumption 1.6 follows from Assumption C.2 with $\beta_i = 1/n$ for all $i \in [n]$.

We now prove the other direction. First, we fix arbitrary vectors $x \in \mathbb{R}^d$ and $u_1, \ldots, u_n \in \mathbb{R}^d$, and weights $\beta_1, \ldots, \beta_n \in \mathbb{R}$ such that $0 \leq \beta_i \leq 1$ for all $i \in [n]$ and $\sum_{i=1}^n \beta_i = 1$. Assume that $\beta := (\beta_1, \ldots, \beta_n) \in \mathbb{Q}^n$. Then there exists a common denominator k such that $\beta = \left(\frac{k_1}{k}, \ldots, \frac{k_n}{k}\right)$ and $\sum_{i=1}^n k_i = k$. Let us define a sequence of vectors $\tilde{u} = (\underbrace{u_1, \ldots, u_1}_{k_1}, \ldots, \underbrace{u_n, \ldots, u_n}_{k_n})$. Then,

$$\left\|\sum_{i=1}^n \beta_i(\nabla f(x+u_i) - \nabla f(x))\right\|^2 = \left\|\frac{1}{k}\sum_{i=1}^n k_i(\nabla f(x+u_i) - \nabla f(x))\right\|^2$$
$$= \left\|\frac{1}{k}\sum_{i=1}^n\sum_{j=1}^{k_i}(\nabla f(x+u_i) - \nabla f(x))\right\|^2 = \left\|\frac{1}{k}\sum_{i=1}^k(\nabla f(x+\tilde{u}_i) - \nabla f(x))\right\|^2.$$

Due to Assumption 1.6 we get

$$\left\|\frac{1}{k}\sum_{i=1}^k(\nabla f(x+\tilde{u}_i) - \nabla f(x))\right\|^2 \leq L_A^2\left(\frac{1}{k}\sum_{i=1}^k \|\tilde{u}_i\|^2\right) + L_B^2\left\|\frac{1}{k}\sum_{i=1}^k \tilde{u}_i\right\|^2$$
$$= L_A^2\left(\frac{1}{k}\sum_{i=1}^n k_i \|u_i\|^2\right) + L_B^2\left\|\frac{1}{k}\sum_{i=1}^n k_i u_i\right\|^2.$$

Since $\beta_i = \frac{k_i}{k}$, the proof of this case is completed. To finish the proof for an arbitrary β satisfying these constraints, it remains to note that the inequality has the form $g(\beta) \leq 0$, where g is a continuous function on this simplex in $\mathbb{R}^n$. Since the inequality holds for all rational points of this simplex, which are dense in it, by continuity, it extends to all points of the simplex. □

D Auxiliary Lemmas and Definitions

Definition D.1. Assume that S is a uniformly random subset from $[d]$, $|S| = K$, $K \in [d]$. A stochastic mapping $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is Rand$K$ if

$$\mathcal{C}(x; S) = \frac{d}{K} \sum_{j \in S} x_j e_j,$$

where $\{e_i\}_{i=1}^d$ is the standard unit basis.

Lemma D.2 (Variance decomposition; Folklore result)**.** *Let $x \in \mathbb{R}^d$ be a random vector with finite mean and variance. Then for any deterministic vector $c \in \mathbb{R}^d$, we have the identity*

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

Lemma D.3. *Let $x_i \in \mathbb{R}^d, i \in [k]$ be mutually independent random vectors with expectations $\mathbb{E}[x_i]$. Then, for all $\beta_1, \ldots, \beta_k \in \mathbb{R}$,*

$$\mathbb{E}\left[\left\|\sum_{i=1}^k \beta_i (x_i - \mathbb{E}[x_i])\right\|^2\right] = \sum_{i=1}^k \beta_i^2 \mathbb{E}\left[\|x_i - \mathbb{E}[x_i]\|^2\right].$$

Proof. It suffices to show that the cross-terms vanish. For any $i, j \in [k]$ with $i \neq j$, by the independence of x_i and x_j, we obtain

$$\mathbb{E}\left[\langle \beta_i (x_i - \mathbb{E}[x_i]), \beta_j (x_j - \mathbb{E}[x_j])\rangle\right] = \beta_i \beta_j \left\langle \mathbb{E}[x_i - \mathbb{E}[x_i]], \mathbb{E}[x_j - \mathbb{E}[x_j]] \right\rangle = 0.$$

□

Lemma D.4. *Let $\mathcal{S}$ be a set of indices. Suppose compressors $\mathcal{C}_i \in \mathbb{U}(\omega)$ for all $i \in \mathcal{S}$ (Definition 1.5), and let $\beta_i \geq 0$ be scalars such that $\sum_{i \in \mathcal{S}} \beta_i = 1$. Then the compressor $\bar{\mathcal{C}}(x) := \sum_{i \in \mathcal{S}} \beta_i \mathcal{C}_i(x)$ belongs to $\mathbb{U}\left(\omega \sum_{i \in \mathcal{S}} \beta_i^2\right)$.*

Proof. First, we show that $\bar{\mathcal{C}}(x)$ is unbiased

$$\mathbb{E}\left[\bar{\mathcal{C}}(x)\right] = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbb{E}\left[\mathcal{C}_i(x)\right] = x.$$

Then, since the compressors $\mathcal{C}_i$ are mutually independent, we may use Lemma D.3 and get

$$\begin{aligned}
&\mathbb{E}\left[\left\|\bar{\mathcal{C}}(x) - x\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\sum_{i \in \mathcal{S}} \beta_i \left(\mathcal{C}_i(x) - x\right)\right\|^2\right] = \sum_{i \in \mathcal{S}} \beta_i^2 \mathbb{E}\left[\left\|\mathcal{C}_i(x) - x\right\|^2\right] \leq \left(\omega \sum_{i \in \mathcal{S}} \beta_i^2\right) \|x\|^2.
\end{aligned}$$

□

Lemma D.5. *Consider $f(\gamma) = \frac{1}{2\gamma} - c - d\gamma$, where $d = \sum_{i=1}^k d_i$ and $c > 0, d_i > 0$ for all $i \in [k]$, then for any $\gamma \in \left(0, \frac{1}{(k+1)\max\left\{2c, \sqrt{2d_1}, \cdots, \sqrt{2d_k}\right\}}\right]$ holds $f(\gamma) \geq 0$.*

Proof. Clearly,

$$f(\gamma) = \frac{1 - 2c\gamma - 2d\gamma^2}{2\gamma} \geq 0 \quad \iff \quad \frac{d\gamma^2 + c\gamma - \frac{1}{2}}{\gamma} \leq 0.$$

We find the zeros of the numerator:

$$D = c^2 + 4d \cdot \frac{1}{2} \quad \iff \quad \gamma = \frac{-c \pm \sqrt{c^2 + 2d}}{2d}.$$

Thus, the maximum γ that satisfies the inequality is

$$\gamma_{\max} = \frac{-c + \sqrt{c^2 + 2d}}{2d} = \frac{(-c + \sqrt{c^2 + 2d})(c + \sqrt{c^2 + 2d})}{2d(c + \sqrt{c^2 + 2d})} = \frac{1}{c + \sqrt{c^2 + 2d}}.$$

and $f(\gamma) \geq 0$ for any $\gamma \in (0, \gamma_{\max}]$. To finish the proof, it remains to note that

$$\gamma_{\max} \geq \frac{1}{2c + \sqrt{2d}} = \frac{1}{2c + \sqrt{2\sum_{i=1}^{k} d_i}} \geq \frac{1}{2c + \sum_{i=1}^{k} \sqrt{2d_i}} \geq \frac{1}{(k+1)\max\left\{2c, \sqrt{2d_1}, \cdots, \sqrt{2d_k}\right\}}.$$

□

Lemma D.6. *Consider the function $g(x) = ax^3 + bx^2 + cx - 1$. If $a, b, c > 0$ then there exists a unique x_0 such that $g(x_0) = 0$ and $x_0 \in \left[\frac{\bar{x}}{2}, \bar{x}\right]$, where $\bar{x} = \frac{1}{\max\left\{\sqrt[3]{a}, \sqrt{b}, c\right\}}$.*

Proof. Since g is strictly increasing on $[0, +\infty)$, there exists at most one nonnegative solution to the equation $g(x) = 0$. Then, we compute g at the points $\bar{x}$ and $\frac{\bar{x}}{2}$. Assume $\sqrt{b} \geq \max\left\{\sqrt[3]{a}, c\right\}$, so $\bar{x} = \frac{1}{\sqrt{b}}$ (other cases are similar).

$$g(\bar{x}) = \frac{a}{b^{\frac{3}{2}}} + \frac{b}{b} + \frac{c}{\sqrt{b}} - 1 > 0 \quad \text{and} \quad g\left(\frac{\bar{x}}{2}\right) = \frac{a}{8b^{\frac{3}{2}}} + \frac{b}{4b} + \frac{c}{2\sqrt{b}} - 1 < 0. \tag{19}$$

We obtain the last inequality since $\sqrt{b} \geq \max\left\{\sqrt[3]{a}, c\right\}$ and $\frac{1}{8} + \frac{1}{4} + \frac{1}{2} < 1$. The inequalities in (19) and the fact that g is strictly increasing on $[0, +\infty)$ imply that $x_0 \in \left[\frac{\bar{x}}{2}, \bar{x}\right]$. □

Lemma D.7. *Assume function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies Assumption 1.1 and Assumption 1.2. Then, we have $\left\|\nabla f(x^0)\right\|^2 \leq 2L\Delta$.*

Proof. Consider the property of L-smoothness:

$$f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2.$$

The right-hand side is a quadratic function. Its minimum is attained at the point

$$y^* = x - \frac{1}{L}\nabla f(x).$$

Substituting $y = y^*$ into the smoothness inequality yields

$$f(y^*) - f(x) \leq -\frac{1}{L}\left\|\nabla f(x)\right\|^2 + \frac{1}{2L}\left\|\nabla f(x)\right\|^2 = -\frac{1}{2L}\left\|\nabla f(x)\right\|^2.$$

Rearranging the terms and setting $x = x^0$, we obtain

$$\left\|\nabla f(x^0)\right\|^2 \leq 2L(f(x^0) - f(y^*)) \leq 2L\Delta.$$

□

Lemma D.8. *Let $w_i > 0$ for all $i = 1, \ldots, n$. Consider the optimization problem*

$$\begin{cases} \sum_{i=1}^{n} \beta_i^2 w_i^{-1} \to \min\limits_{\beta} \\ \sum_{i=1}^{n} \beta_i = 1, \quad 0 \leq \beta_i \leq 1. \end{cases}$$

The unique optimal solution is given by

$$\beta_i^* = \frac{w_i}{\sum_{j=1}^{n} w_j},$$

and the corresponding minimum value equals $\left(\sum_{j=1}^{n} w_j\right)^{-1}$.

Proof. We first solve the problem subject only to the equality constraint $\sum_{i=1}^n \beta_i = 1$ using the method of Lagrange multipliers. Define the Lagrangian

$$\mathcal{L}(\beta, \lambda) = \sum_{i=1}^n \frac{\beta_i^2}{w_i} - \lambda \left(\sum_{i=1}^n \beta_i - 1 \right).$$

Setting the partial derivatives to zero yields

$$\frac{\partial \mathcal{L}}{\partial \beta_i} = \frac{2\beta_i}{w_i} - \lambda = 0 \quad \Rightarrow \quad \beta_i = \frac{\lambda w_i}{2}, \quad i = 1, \ldots, n.$$

Substituting this expression into the equality constraint gives $\frac{\lambda}{2} \sum_{j=1}^n w_j = 1$, hence $\frac{\lambda}{2} = \left(\sum_{j=1}^n w_j \right)^{-1}$. This immediately implies

$$\beta_i^* = \frac{w_i}{\sum_{j=1}^n w_j}.$$

Since $w_i > 0$, we have $\beta_i^* > 0$. Furthermore, $\beta_i^* \leq 1$ because $w_i \leq \sum_{j=1}^n w_j$ for all i. Thus, the box constraints $0 \leq \beta_i \leq 1$ are automatically satisfied. Finally, evaluating the objective at β^* gives

$$\sum_{i=1}^n \frac{(\beta_i^*)^2}{w_i} = \sum_{i=1}^n \frac{w_i^2}{w_i \left(\sum_{j=1}^n w_j \right)^2} = \frac{\sum_{i=1}^n w_i}{\left(\sum_{j=1}^n w_j \right)^2} = \left(\sum_{j=1}^n w_j \right)^{-1},$$

which completes the proof. □

Lemma D.9. *Suppose that Assumption 1.1 holds and let $x^{k+1} = x^k - \gamma g^k$. Then for any $g^k \in \mathbb{R}^d$ and $\gamma > 0$, we have*

$$f(x^{k+1}) \leq f(x^k) - \frac{\gamma}{2} \left\| \nabla f(x^k) \right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2} \right) \left\| x^{k+1} - x^k \right\|^2 + \frac{\gamma}{2} \left\| g^k - \nabla f(x^k) \right\|^2. \quad (20)$$

Proof. Using $L-$smoothness, we have

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \left\langle \nabla f(x^k), x^{k+1} - x^k \right\rangle + \frac{L}{2} \left\| x^{k+1} - x^k \right\|^2 \\ &= f(x^k) - \gamma \left\langle \nabla f(x^k), g^k \right\rangle + \frac{L}{2} \left\| x^{k+1} - x^k \right\|^2. \end{aligned}$$

Next, due to $-\langle x, y \rangle = \frac{1}{2} \|x - y\|^2 - \frac{1}{2} \|x\|^2 - \frac{1}{2} \|y\|^2$, we obtain

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

□

E Proofs for Heterogeneous-Time Inkheart SGD and Inkheart SGD

We provide proofs for Heterogeneous-Time Inkheart SGD. Then, the results for Inkheart SGD are corollaries, since Heterogeneous-Time Inkheart SGD is a generalization of Inkheart SGD.

We denote by $\mathbb{E}_k[\cdot]$ the expectation conditional on all workers receiving x_i^k and by $\mathbb{E}_{k,\xi}[\cdot]$ the expectation conditional on all workers computing the stochastic gradients at iteration k.

Theorem E.1. *Let Assumptions 1.1, 1.2, 1.3 and 1.6 be satisfied and suppose that $\mathcal{C}_{ij}$ satisfies Definition 1.5 with parameter ω, $\mathcal{C}_{s,ij}$ satisfies Definition 1.5 with parameter ω_s. Consider Heterogeneous-Time Inkheart SGD with arbitrary $m_i, b_i, \ell_i > 0$ and weights β_i (not necessarily defined as in (17); it is sufficient to assume that $\beta_i \in [0,1]$ and $\sum_{i=1}^n \beta_i = 1$) chosen to satisfy*

$$\begin{cases} 8 \sum\limits_{i=1}^n \beta_i^2 \frac{\omega}{m_i} \leq 1 \\ 8 \sum\limits_{i=1}^n \beta_i^2 \left(\frac{\omega \sigma^2}{m_i b_i \varepsilon} + \frac{\sigma^2}{b_i \varepsilon} \right) \leq 1. \end{cases} \quad (21)$$

Then, Heterogeneous-Time Inkheart SGD with

$$\gamma = \frac{1}{6} \times \min\left\{ \frac{1}{L_{\max}}, \frac{1}{L_{\max}\sqrt{\sum\limits_{i=1}^{n}\left(\frac{\omega\omega_{\rm s}}{pm_i\ell_i} + \frac{\omega_{\rm s}}{p\ell_i}\right)\beta_i^2}}, \frac{1}{L_A\sqrt{\frac{1}{p}\sum\limits_{i=1}^{n}\frac{\omega_{\rm s}\beta_i}{\ell_i}}}, \right\} \tag{22}$$

converges after at most

$$K = 48 \times \frac{\Delta}{\varepsilon} \max\left\{ L_{\max}, L_{\max}\sqrt{\sum_{i=1}^{n}\left(\frac{\omega\omega_{\rm s}}{pm_i\ell_i} + \frac{\omega_{\rm s}}{p\ell_i}\right)\beta_i^2}, L_A\sqrt{\frac{1}{p}\sum_{i=1}^{n}\frac{\omega_{\rm s}\beta_i}{\ell_i}} \right\} \tag{23}$$

iterations.

Proof. Using Assumption 1.1 and Lemma D.9, we get

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

Then, we take conditional expectation from both parts and use Lemma E.7 to bound the last term

$$\begin{aligned}\mathbb{E}_k\left[f(x^{k+1})\right] \leq f(x^k) &- \frac{\gamma}{2}\left\|\nabla f(x^k)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\mathbb{E}_k\left[\left\|x^{k+1} - x^k\right\|^2\right] \\ &+ 2\gamma\left(\sum_{i=1}^{n}\beta_i^2\frac{\omega}{m_i}\right)\left\|\nabla f(x^k)\right\|^2 + \gamma\sum_{i=1}^{n}\left(L_A^2\beta_i + 2L^2\beta_i^2\frac{\omega}{m_i}\right)\left\|x_i^k - x^k\right\|^2 \\ &+ \gamma L_B^2\left\|\sum_{i=1}^{n}\beta_i\left(x_i^k - x^k\right)\right\|^2 + \gamma\sum_{i=1}^{n}\beta_i^2\left(\frac{\omega\sigma^2}{m_ib_i} + \frac{\sigma^2}{b_i}\right),\end{aligned} \tag{24}$$

By Assumption of the theorem, the inequalities $2\gamma\left(\sum\limits_{i=1}^{n}\beta_i^2\frac{\omega}{m_i}\right)\left\|\nabla f(x^k)\right\|^2 \leq \frac{\gamma}{4}\left\|\nabla f(x^k)\right\|^2$ and $\gamma\sum\limits_{i=1}^{n}\beta_i^2\left(\frac{\omega\sigma^2}{m_ib_i} + \frac{\sigma^2}{b_i}\right) \leq \frac{\gamma\varepsilon}{8}$ hold. Then, we sum the inequalities (36) and (37) from Lemma E.9 multiplied by $\sum\limits_{i=1}^{n}\kappa_i$ and η, respectively:

$$\begin{aligned}&\sum_{i=1}^{n}\kappa_i\beta_i\mathbb{E}_k\left[\left\|x_i^{k+1} - x^{k+1}\right\|^2\right] + \eta\mathbb{E}_k\left[\left\|\sum_{i=1}^{n}\beta_i\left(x_i^{k+1} - x^{k+1}\right)\right\|^2\right] \\ &\leq (1-p)\left[\sum_{i=1}^{n}\frac{\omega_{\rm s}\kappa_i\beta_i}{\ell_i}\mathbb{E}_k\left[\left\|x^{k+1} - x^k\right\|^2\right] + \sum_{i=1}^{n}\beta_i\kappa_i\left\|x_i^k - x^k\right\|^2\right] \\ &\quad + (1-p)\left[\sum_{i=1}^{n}\frac{\omega_{\rm s}\eta\beta_i^2}{\ell_i}\mathbb{E}_k\left[\left\|x^{k+1} - x^k\right\|^2\right] + \eta\left\|\sum_{i=1}^{n}\beta_i\left(x_i^k - x^k\right)\right\|^2\right].\end{aligned} \tag{25}$$

Then, we add the inequality (25) to both parts of the inequality (24) to construct a Lyapunov function.

$$\begin{aligned}&\mathbb{E}_k\left[f(x^{k+1})\right] + \sum_{i=1}^{n}\kappa_i\beta_i\mathbb{E}_k\left[\left\|x_i^{k+1} - x^{k+1}\right\|^2\right] + \eta\mathbb{E}_k\left[\left\|\sum_{i=1}^{n}\beta_i\left(x_i^{k+1} - x^{k+1}\right)\right\|^2\right] \\ &\leq f(x^k) - \frac{\gamma}{4}\left\|\nabla f(x^k)\right\|^2 \\ &- \underbrace{\left(\frac{1}{2\gamma} - \frac{L}{2} - (1-p)\sum_{i=1}^{n}\frac{\omega_{\rm s}\kappa_i\beta_i}{\ell_i} - (1-p)\sum_{i=1}^{n}\frac{\omega_{\rm s}\eta\beta_i^2}{\ell_i}\right)}_{A}\mathbb{E}_k\left[\left\|x^{k+1} - x^k\right\|^2\right]\end{aligned} \tag{26}$$

$$+\sum_{i=1}^{n}\underbrace{\left(\gamma L_A^2\beta_i+2\gamma L^2\beta_i^2\frac{\omega}{m_i}+(1-p)\kappa_i\beta_i\right)}_{B_i}\left\|x_i^k-x^k\right\|^2$$

$$+\underbrace{\left(\gamma L_B^2+(1-p)\eta\right)}_{C}\left\|\sum_{i=1}^{n}\beta_i\left(x_i^k-x^k\right)\right\|^2+\frac{\gamma\varepsilon}{8}.$$

Next, we should choose κ_i, η to achieve $\eta = C$ and $\kappa_i\beta_i = B_i$ for all $i \in [n]$:

$$\begin{cases}\kappa_i = 2\gamma L^2\beta_i\frac{\omega}{m_i}+\gamma L_A^2+\kappa_i(1-p)\\ \eta=\left(\gamma L_B^2+\eta(1-p)\right)\end{cases}\qquad\qquad\begin{cases}\kappa_i=\frac{2\gamma\omega L^2\beta_i}{pm_i}+\frac{\gamma L_A^2}{p}\\ \eta=\frac{\gamma L_B^2}{p}\end{cases}.$$

Then, we substitute $\kappa_i, i \in [n]$ and η into A:

$$\begin{aligned}A&=\frac{1}{2\gamma}-\frac{L}{2}-(1-p)\sum_{i=1}^{n}\frac{\omega_{\mathrm{s}}\kappa_i\beta_i}{\ell_i}-(1-p)\sum_{i=1}^{n}\frac{\omega_{\mathrm{s}}\eta\beta_i^2}{\ell_i}\\&\geq\frac{1}{2\gamma}-\frac{L}{2}-\sum_{i=1}^{n}\frac{\omega_{\mathrm{s}}\kappa_i\beta_i}{\ell_i}-\sum_{i=1}^{n}\frac{\omega_{\mathrm{s}}\eta\beta_i^2}{\ell_i}\\&=\frac{1}{2\gamma}-\frac{L}{2}-\gamma\frac{2\omega\omega_{\mathrm{s}}L^2}{p}\sum_{i=1}^{n}\frac{\beta_i^2}{\ell_i m_i}-\gamma\frac{L_A^2}{p}\sum_{i=1}^{n}\frac{\omega_{\mathrm{s}}\beta_i}{\ell_i}-\gamma\frac{L_B^2}{p}\sum_{i=1}^{n}\frac{\omega_{\mathrm{s}}\beta_i^2}{\ell_i}\\&\geq\frac{1}{2\gamma}-L-2\gamma\left(\underbrace{\frac{\omega\omega_{\mathrm{s}}L^2}{p}\sum_{i=1}^{n}\frac{\beta_i^2}{\ell_i m_i}+\frac{L_B^2}{p}\sum_{i=1}^{n}\frac{\omega_{\mathrm{s}}\beta_i^2}{\ell_i}}_{d_1}+\underbrace{\frac{L_A^2}{p}\sum_{i=1}^{n}\frac{\omega_{\mathrm{s}}\beta_i}{\ell_i}}_{d_2}\right).\end{aligned}$$

Applying Lemma D.5 to the function $g(\gamma)=\frac{1}{2\gamma}-L-(2d_1+2d_2)\gamma$, we conclude that $A\geq 0$ holds whenever

$$\gamma\leq\frac{1}{3}\times\frac{1}{\max\left\{2L,\sqrt{2\cdot 2d_1},\sqrt{2\cdot 2d_2}\right\}}=\frac{1}{6}\times\frac{1}{\max\left\{L,\sqrt{d_1},\sqrt{d_2}\right\}}.$$

By definition, $\max\{L, L_B\} \leq L_{\max}$, which implies

$$d_1\leq L_{\max}^2\left(\frac{\omega\omega_{\mathrm{s}}}{p}\sum_{i=1}^{n}\frac{\beta_i^2}{\ell_i m_i}+\frac{1}{p}\sum_{i=1}^{n}\frac{\omega_{\mathrm{s}}\beta_i^2}{\ell_i}\right).$$

Thus, our choice of γ in (22) guarantees $A \geq 0$. Dropping the non-positive term $-A\cdot\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right]$ and using $\eta = C$, $\kappa_i\beta_i = B_i$, we transform the inequality (26) into

$$\mathbb{E}_k\left[f(x^{k+1})\right]+\sum_{i=1}^{n}\kappa_i\beta_i\mathbb{E}_k\left[\left\|x_i^{k+1}-x^{k+1}\right\|^2\right]+\eta\mathbb{E}_k\left[\left\|\sum_{i=1}^{n}\beta_i\left(x_i^{k+1}-x^{k+1}\right)\right\|^2\right]$$
$$\leq f(x^k)-\frac{\gamma}{4}\left\|\nabla f(x^k)\right\|^2+\sum_{i=1}^{n}\kappa_i\beta_i\left\|x_i^k-x^k\right\|^2+\eta\left\|\sum_{i=1}^{n}\beta_i\left(x_i^k-x^k\right)\right\|^2+\frac{\gamma\varepsilon}{8}.$$

Taking the full expectation and summing for $k \in [0, \ldots, K-1]$, we get

$$\mathbb{E}\left[f(x^K)\right]+\sum_{i=1}^{n}\kappa_i\beta_i\mathbb{E}\left[\left\|x_i^K-x^K\right\|^2\right]+\eta\mathbb{E}\left[\left\|\sum_{i=1}^{n}\beta_i\left(x_i^K-x^K\right)\right\|^2\right]$$

$$\leq f(x^0) - \frac{\gamma}{4}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(x^k)\right\|^2\right] + \sum_{i=1}^{n}\kappa_i\beta_i\left\|x_i^0 - x^0\right\|^2 + \eta\left\|\sum_{i=1}^{n}\beta_i\left(x_i^0 - x^0\right)\right\|^2 + K\frac{\gamma\varepsilon}{8}.$$

From the construction of Heterogeneous-Time Inkheart SGD, $x_i^0 = x^0$ for $i \in [n]$. Ignoring non-negative terms we finally obtain

$$\frac{\gamma}{4}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(x^k)\right\|^2\right] \leq f(x^0) - \mathbb{E}\left[f(x^K)\right] + K\frac{\gamma\varepsilon}{8} \quad \Longrightarrow \quad \sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(x^k)\right\|^2\right] \leq \frac{4\Delta}{\gamma} + K\frac{\varepsilon}{2}.$$

To achieve $\frac{1}{K}\sum\limits_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(x^k)\right\|^2\right] \leq \varepsilon$ we should take $\frac{4\Delta}{\gamma K} \leq \frac{\varepsilon}{2}$. Thus, it is sufficient to run Heterogeneous-Time Inkheart SGD for $K = \frac{8\Delta}{\gamma\varepsilon}$ iterations. □

E.1 Time Complexity

In this section, we restrict our attention to the case where all compressors used by the server and the workers are Rand1. We begin by selecting the probability p. The following lemma shows which probability p we can choose.

Lemma E.2. *Consider the probabilistic mechanism from Heterogeneous-Time Inkheart SGD and assume that the compressors $\mathcal{C}_{s,ij}$ are Rand1 (which means the server sends only one coordinate at a time). Then, for $p = \min\left\{\frac{\ell_{\min}}{d}, 1\right\}$, where $\ell_{\min} := \left\lfloor\frac{t}{\kappa_{\max}}\right\rfloor$ and $\kappa_{\max} := \max\limits_{i\in[n]}\kappa_i$, the condition $\mathbb{E}\left[t_{\text{server}}^k\right] \leq 2t$ holds.*

Proof. Sending one coordinate to worker i takes at most κ_i seconds and the server either sends the full point (d coordinates) or sends a compressed point, which takes less than t seconds by the definition of the time budget. When the server sends the full point, the workers synchronize and wait for the slowest worker. Thus, we obtain

$$\mathbb{E}\left[t_{\text{server}}^k\right] \leq p\kappa_{\max}d + (1-p)t \leq p\kappa_{\max}d + t$$

Since $p = \min\left\{\frac{\ell_{\min}}{d}, 1\right\} \leq \frac{\ell_{\min}}{d}$, we get

$$\mathbb{E}\left[t_{\text{server}}^k\right] \leq \frac{\ell_{\min}}{d}\kappa_{\max}d + t = \left\lfloor\frac{t}{\kappa_{\max}}\right\rfloor\kappa_{\max} + t \leq 2t.$$

□

Below we obtain the optimal time budget t, the optimal weights w_i, and the time complexity for the scenario when all workers have different time performances h_i, τ_i and κ_i for $i \in [n]$.

Theorem 4.2 (Time Complexity of Heterogeneous-Time Inkheart SGD)**.** *Let Assumptions 1.1, 1.2, 1.3, and 1.6 be satisfied. Additionally, assume that Assumption 4.1 is satisfied and the workers and the server use Rand1 compressors. Let $b_i = \left\lfloor\frac{t}{h_i}\right\rfloor$, $m_i = \left\lfloor\frac{t}{\tau_i}\right\rfloor$, $\ell_i = \left\lfloor\frac{t}{\kappa_i}\right\rfloor$, where $\ell_{\min} = \left\lfloor\frac{t}{\kappa_{\max}}\right\rfloor$, $t = \max\left\{\max_{i\in[n]}\left\{h_i, \tau_i, \kappa_i\right\}, s^*\right\}$, $\kappa_{\max} = \max_{i\in[n]}\kappa_i$, and s^* is the solution of the equation*

$$\left(\sum_{i=1}^{n}\frac{1}{\frac{16\omega\tau_i}{s} + \frac{16\sigma^2h_i}{\varepsilon s} + \frac{32\sigma^2\omega h_i\tau_i}{\varepsilon s^2} + \frac{4d\omega_s\kappa_{\max}\kappa_i}{s^2} + \frac{8d\omega_s\omega\kappa_{\max}\kappa_i\tau_i}{s^3}}\right)^{-1} = 1. \tag{15}$$

Then, the expected time for (Heterogeneous-Time Inkheart SGD) with $p = \min\left\{\frac{\ell_{\min}}{d}, 1\right\}$ to find an ε–stationary point of (1) is at most

$$\mathcal{O}\left(\frac{tL_{\max}\Delta}{\varepsilon} + \frac{d\kappa_{\max}L_A\Delta}{\varepsilon}\right). \tag{16}$$

Proof. First, suppose that $p < 1$. We require $t \geq \max\limits_{i\in[n]} \{h_i, \tau_i, \kappa_i\}$ to ensure that at least one gradient is computed and compressed by each worker, and the server sends a new point compressed by at least one compressor. Then, we show that with our choice of t the conditions (21) are satisfied and Heterogeneous-Time Inkheart SGD converges. Notice that

$$\begin{cases} 8\sum\limits_{i=1}^{n} \beta_i^2 \frac{\omega}{m_i} \leq \sum\limits_{i=1}^{n} \beta_i^2 w_i^{-1} \leq 1 \\ 8\sum\limits_{i=1}^{n} \beta_i^2 \left(\frac{\omega\sigma^2}{m_i b_i \varepsilon} + \frac{\sigma^2}{b_i\varepsilon}\right) \leq \sum\limits_{i=1}^{n} \beta_i^2 w_i^{-1} \leq 1. \end{cases}$$

The last inequalities follow from Lemma E.3. Moreover, using Lemma E.3, and our choice of β_i and t, we have $\sum\limits_{i=1}^{n} \left(\frac{\omega\omega_{\mathrm{s}}}{pm_i\ell_i} + \frac{\omega_{\mathrm{s}}}{p\ell_i}\right)\beta_i^2 \leq \sum\limits_{i=1}^{n} \beta_i^2 w_i^{-1} \leq 1$. Therefore,

$$L_{\max}\sqrt{\sum_{i=1}^{n}\left(\frac{\omega\omega_{\mathrm{s}}}{pm_i\ell_i} + \frac{\omega_{\mathrm{s}}}{p\ell_i}\right)\beta_i^2} \leq L_{\max}. \tag{27}$$

Recall that the time of each iteration consists of computing gradients, sending compressed vectors by the workers and sending compressed points by the server. Using Lemma E.2, we bound the time complexity as $\mathbb{E}\left[T_{\text{time}}\right] \leq 4t \times K$, where K is the iteration complexity (23) from Theorem E.1:

$$\mathbb{E}\left[T_{\text{time}}\right] \leq 192t \times \frac{\Delta}{\varepsilon}\max\left\{L_{\max}, L_{\max}\sqrt{\sum_{i=1}^{n}\left(\frac{\omega\omega_{\mathrm{s}}}{pm_i\ell_i} + \frac{\omega_{\mathrm{s}}}{p\ell_i}\right)\beta_i^2}, L_A\sqrt{\frac{1}{p}\sum_{i=1}^{n}\frac{\omega_{\mathrm{s}}\beta_i}{\ell_i}}\right\}$$

Then, using (27), we obtain

$$\mathbb{E}\left[T_{\text{time}}\right] \leq 192 \times \frac{\Delta}{\varepsilon}\max\left\{L_{\max}t, t \times L_A\sqrt{\frac{1}{p}\sum_{i=1}^{n}\frac{\omega_{\mathrm{s}}\beta_i}{\ell_i}}\right\}. \tag{28}$$

To finish the proof we consider the last term from (28) and use bounds $\frac{1}{\ell_i} \leq \frac{2\kappa_i}{t}$ and $\frac{1}{p} = \frac{d}{\ell_{\min}} \leq \frac{d}{\frac{t}{2\kappa_{\max}}} = \frac{2d\kappa_{\max}}{t}$. Finally, we obtain

$$t \times L_A\sqrt{\frac{1}{p}\sum_{i=1}^{n}\beta_i\frac{\omega_{\mathrm{s}}}{\ell_i}} \leq t \times L_A\sqrt{\frac{2d\kappa_{\max}}{t}\sum_{i=1}^{n}\beta_i\frac{2\omega_{\mathrm{s}}\kappa_i}{t}} = 2L_A\sqrt{\kappa_{\max}d\omega_{\mathrm{s}}\sum_{i=1}^{n}\beta_i\kappa_i} \leq 2dL_A\kappa_{\max}.$$

Here we used $\omega_{\mathrm{s}} \leq d-1$ and $\sum_{i=1}^{n}\beta_i = 1$. Now, consider the case $p = 1$. Then, $x_i^k = x^k$ for all $i \in [n]$ and $k \geq 0$, and the server compressors are not used, so they can be replaced by the identity compressors with $\omega_{\mathrm{s}} = 0$. Similarly to Lemma E.3 using (15), $t \geq s^*$, and the bounds $b_i \geq t/(2h_i)$ and $m_i \geq t/(2\tau_i)$, we obtain $\sum_{i=1}^{n} w_i \geq 1$. Therefore, $\sum_{i=1}^{n}\beta_i^2 w_i^{-1} = \left(\sum_{i=1}^{n} w_i\right)^{-1} \leq 1$, which implies (21). Theorem E.1 with $\omega_{\mathrm{s}} = 0$ yields $\gamma = 1/(6L_{\max})$ and $K = 48L_{\max}\Delta/\varepsilon$. By Lemma E.2, $\mathbb{E}\left[T_{\text{time}}\right] \leq 4tK = \mathcal{O}\left(\frac{tL_{\max}\Delta}{\varepsilon}\right)$. □

Lemma E.3. *Consider the time budget t defined in Theorem 4.2 and suppose that $p < 1$. Then, the following inequality holds*

$$\sum_{i=1}^{n}\beta_i^2\underbrace{\left(\frac{8\omega}{m_i} + \frac{8\sigma^2}{\varepsilon b_i} + \frac{8\sigma^2\omega}{\varepsilon b_i m_i} + \frac{\omega_{\mathrm{s}}}{p\ell_i} + \frac{d\omega_{\mathrm{s}}\omega}{\ell_{\min}\ell_i m_i}\right)}_{w_i^{-1}} \leq 1.$$

Proof. With our choice of β_i defined in (17), we obtain

$$\sum_{i=1}^{n}\beta_i^2 w_i^{-1} = \frac{1}{\sum\limits_{j=1}^{n} w_j} = \left(\sum_{j=1}^{n} w_j\right)^{-1}. \tag{29}$$

Since $b_i = \left\lfloor \frac{t}{h_i} \right\rfloor \geq \frac{t}{2h_i}$, $m_i = \left\lfloor \frac{t}{\tau_i} \right\rfloor \geq \frac{t}{2\tau_i}$ and $\ell_i = \left\lfloor \frac{t}{\kappa_i} \right\rfloor \geq \frac{t}{2\kappa_i}$, we get

$$
\begin{aligned}
w_i &= \left(\frac{8\omega}{m_i} + \frac{8\sigma^2\omega}{\varepsilon b_i m_i} + \frac{8\sigma^2}{\varepsilon b_i} + \frac{d\omega_{\mathrm{s}}}{\ell_{\min}\ell_i} + \frac{d\omega_{\mathrm{s}}\omega}{\ell_{\min}\ell_i m_i} \right)^{-1} \\
&\geq \left(\frac{8\omega}{\frac{t}{2\tau_i}} + \frac{8\sigma^2\omega}{\varepsilon \frac{t}{2h_i}\frac{t}{2\tau_i}} + \frac{8\sigma^2}{\varepsilon \frac{t}{2h_i}} + \frac{d\omega_{\mathrm{s}}}{\frac{t}{2\kappa_{\max}}\frac{t}{2\kappa_i}} + \frac{d\omega_{\mathrm{s}}\omega}{\frac{t}{2\kappa_{\max}}\frac{t}{2\kappa_i}\frac{t}{2\tau_i}} \right)^{-1} \\
&= \left(\frac{16\omega\tau_i}{t} + \frac{16\sigma^2 h_i}{\varepsilon t} + \frac{32\sigma^2\omega h_i\tau_i}{\varepsilon t^2} + \frac{4d\omega_{\mathrm{s}}\kappa_{\max}\kappa_i}{t^2} + \frac{8d\omega_{\mathrm{s}}\omega\kappa_{\max}\kappa_i\tau_i}{t^3} \right)^{-1}. \qquad (30)
\end{aligned}
$$

Substituting (30) into (29) we obtain

$$
\sum_{i=1}^n \beta_i^2 w_i^{-1} \leq \underbrace{\left(\sum_{i=1}^n \frac{1}{\frac{16\omega\tau_i}{t} + \frac{16\sigma^2 h_i}{\varepsilon t} + \frac{32\sigma^2\omega h_i\tau_i}{\varepsilon t^2} + \frac{4d\omega_{\mathrm{s}}\kappa_{\max}\kappa_i}{t^2} + \frac{8d\omega_{\mathrm{s}}\omega\kappa_{\max}\kappa_i\tau_i}{t^3}} \right)^{-1}}_{\delta(t)}.
$$

Recall that $\delta(t)$ is a strictly decreasing function of t and by the definition of the time budget $t \geq s^*$. Since $\delta(s^*) = 1$ the last inequality yields $\sum\limits_{i=1}^n \beta_i^2 w_i^{-1} \leq 1$. □

Theorem 2.2 (Time Complexity of Inkheart SGD)**.** *Consider the assumptions and result of Theorem 2.1. Additionally, assume that Assumption 1.4 is satisfied and the workers and the server use Rand1 compressor (Definition 1.5). Let $b = \left\lfloor \frac{t}{h} \right\rfloor, m = \left\lfloor \frac{t}{\tau} \right\rfloor$, and $\ell = \left\lfloor \frac{t}{\kappa} \right\rfloor$, where $t = \max\left\{ h, \tau, \kappa, {}^{16\omega\tau}/_{n}, {}^{16\sigma^2 h}/_{n\varepsilon}, {}^{2d\kappa}/_{\sqrt{n}}, \sqrt{{}^{32d\sigma^2 h\tau}/_{n\varepsilon}}, \left({}^{8d^3\tau\kappa^2}/_{n}\right)^{\frac{1}{3}} \right\}$. Then, the expected time for Inkheart SGD with $p = \min\left\{ \frac{\ell}{d}, 1 \right\}$ to find an ε–stationary point of (1) is*

$$
\mathcal{O}\left(\max\left\{ h, \tau, \kappa, \frac{d\kappa}{\sqrt{n}}, \frac{d\tau}{n}, \left(\frac{d^3\tau\kappa^2}{n}\right)^{\frac{1}{3}}, \sqrt{\frac{d\sigma^2 h\tau}{n\varepsilon}}, \frac{\sigma^2 h}{n\varepsilon} \right\} \frac{L_{\max}\Delta}{\varepsilon} + \frac{d\kappa L_A \Delta}{\varepsilon} \right). \qquad (11)
$$

Proof. Since Heterogeneous-Time Inkheart SGD generalizes Inkheart SGD, we can use Theorem 4.2. Recall the definition of s^* in Theorem 4.2:

$$
\left(\sum_{i=1}^n \frac{1}{\frac{16\omega\tau_i}{s} + \frac{16\sigma^2 h_i}{\varepsilon s} + \frac{32\sigma^2\omega h_i\tau_i}{\varepsilon s^2} + \frac{4d\omega_{\mathrm{s}}\kappa_{\max}\kappa_i}{s^2} + \frac{8d\omega_{\mathrm{s}}\omega\kappa_{\max}\kappa_i\tau_i}{s^3}} \right)^{-1} = 1.
$$

We rewrite this equation to the case when all the workers are identical.

$$
\begin{aligned}
&\left(\frac{1}{\frac{16\omega\tau}{ns} + \frac{16\sigma^2 h}{\varepsilon ns} + \frac{32\sigma^2\omega h\tau}{\varepsilon ns^2} + \frac{4d\omega_{\mathrm{s}}\kappa^2}{ns^2} + \frac{8d\omega_{\mathrm{s}}\omega\kappa^2\tau}{ns^3}} \right)^{-1} = 1 \iff \\
&\frac{16\omega\tau}{ns} + \frac{16\sigma^2 h}{\varepsilon ns} + \frac{32\sigma^2\omega h\tau}{\varepsilon ns^2} + \frac{4d\omega_{\mathrm{s}}\kappa^2}{ns^2} + \frac{8d\omega_{\mathrm{s}}\omega\kappa^2\tau}{ns^3} = 1.
\end{aligned}
$$

Then, we substitute $r = \frac{1}{s}$. Lemma D.6 yields

$$
s^* \leq \mathcal{O}\left(\max\left\{ \frac{16\omega\tau}{n}, \frac{16\sigma^2 h}{n\varepsilon}, \sqrt{\frac{32\omega\sigma^2 h\tau}{n\varepsilon}}, \sqrt{\frac{4d\omega_{\mathrm{s}}\kappa^2}{n}}, \left(\frac{8d\omega_{\mathrm{s}}\omega\kappa^2\tau}{n}\right)^{\frac{1}{3}} \right\} \right).
$$

Since we use Rand1 compressors, $\omega = \omega_{\mathrm{s}} = d - 1$ and

$$
s^* \leq \mathcal{O}\left(\max\left\{ \frac{16\omega\tau}{n}, \frac{16\sigma^2 h}{n\varepsilon}, \frac{2d\kappa}{\sqrt{n}}, \sqrt{\frac{32d\sigma^2 h\tau}{n\varepsilon}}, \left(\frac{8d^3\tau\kappa^2}{n}\right)^{\frac{1}{3}} \right\} \right).
$$

Applying Theorem 4.2 completes the proof. □

Remark E.4. Let us compare our results with those obtained in (Tyurin et al., 2024). Consider Corollary A.3 which bounds the time complexity $T_{\text{time}} \leq \frac{768L\Delta}{\alpha\varepsilon} \times (\kappa + 2t^*)$ and Example 6.5 which bounds the optimal time budget $t^* \leq \max\left\{h, \tau, \frac{\tau\omega}{n}, \frac{h\sigma^2}{n\varepsilon}, \sqrt{\frac{\tau h\sigma^2\omega}{n\varepsilon}}\right\}$. Note that the previous work focuses on biased server's compressors. If we convert an unbiased Rand1 compressor $\mathcal{C}(x)$ with $\omega = d-1$ into a biased compressor by scaling it as $\frac{1}{d}\mathcal{C}(x)$, we obtain a biased compressor with $\alpha = 1 - \frac{d-1}{d} = \frac{1}{d}$. Thus, for the method described in (Tyurin et al., 2024) the following upper bound holds

$$T_{\text{time}} \leq \frac{768Ld\Delta}{\varepsilon}\left(\kappa + 2 \cdot \max\left\{h, \tau, \frac{\tau d}{n}, \frac{h\sigma^2}{n\varepsilon}, \sqrt{\frac{\tau h\sigma^2 d}{n\varepsilon}}\right\}\right)$$

Note that the time complexity of their method contains terms involving d that do not scale with n.

E.2 Optimal subset of the workers

Previously, we utilized all available workers in Heterogeneous-Time Inkheart SGD. Below, we demonstrate that alternative strategies may achieve better performance. Recall the time bound we obtained in Theorem 4.2. As we observe, it depends on $\kappa_{\max}$ and $\max_{i\in[n]}\{h_i, \tau_i, \kappa_i\}$. Therefore, when we add a new worker, the time complexity does not necessarily decrease. This may happen even if the parameters of the ith worker are relatively small but its κ_i parameter is greater than the $\{\kappa_j\}$ of the other workers. Thus, we should find a subset of workers that yields the optimal time complexity. Algorithm 1 provides a solution. We denote $M_i := \max\{h_i, \tau_i, \kappa_i\}$. We compare $\tilde{T}(S)/L_{\max} = \max\{t(S), d\kappa_{\max}(S)L_A/L_{\max}\}$, which has the same minimizers as $\tilde{T}(S)$ for $L_{\max} > 0$.

Algorithm 1 Optimal Worker Subset Selection

Input: Parameters $\{(h_i, \tau_i, \kappa_i)\}_{i=1}^n, \sigma^2/\varepsilon, \omega, \omega_s, d$, and $L_A/L_{\max} < \infty$ with $L_{\max} > 0$.
Compute $M_i \leftarrow \max\{h_i, \tau_i, \kappa_i\}$ for all $i \in [n]$.
Sort indices by κ in non-decreasing order; let the permutation be π, so $\kappa_{\pi(1)} \leq \cdots \leq \kappa_{\pi(n)}$.
$T_{\min} \leftarrow +\infty, \quad S^* \leftarrow \emptyset$
for $k = 1$ **to** n **do**
  Let $\tilde{S} \leftarrow \{\pi(1), \ldots, \pi(k)\}$.
  Sort workers in $\tilde{S}$ by M_i in non-decreasing order; let the ordering be $\sigma^{(k)}(1), \ldots, \sigma^{(k)}(k)$.
  for $m = 1$ **to** k **do**
    $\hat{S} \leftarrow \{\sigma^{(k)}(1), \ldots, \sigma^{(k)}(m)\}$
    Evaluate $\tilde{T}(\hat{S})/L_{\max}$ using Theorem 4.2.
    if $\tilde{T}(\hat{S})/L_{\max} < T_{\min}$ **then**
      $T_{\min} \leftarrow \tilde{T}(\hat{S})/L_{\max}$
      $S^* \leftarrow \hat{S}$
    end if
  end for
end for
return S^*

Definition E.5. Let $[n]$ be a set of workers' indices. Similarly to Theorem 4.2, we consider

$$t(S) := \max\left\{M_{\max}(S), s^*(S)\right\}, \tag{31}$$

for an arbitrary $\emptyset \neq S \subseteq [n]$, where $M_{\max}(S) := \max_{i\in S} M_i$ and $s^*(S)$ is the solution to the equation

$$\delta_S(s) := \left(\underbrace{\sum_{i\in S} \frac{1}{\frac{16\omega\tau_i}{s} + \frac{16\sigma^2 h_i}{\varepsilon s} + \frac{32\sigma^2\omega h_i\tau_i}{\varepsilon s^2} + \frac{4d\omega_s\kappa_{\max}(S)\kappa_i}{s^2} + \frac{8d\omega_s\omega\kappa_{\max}(S)\kappa_i\tau_i}{s^3}}}_{\psi_S(s)}\right)^{-1} = 1. \tag{32}$$

Here we denote $\kappa_{\max}(S) := \max_{i \in S} \kappa_i$. Theorem 4.2 yields the time complexity when Heterogeneous-Time Inkheart SGD is executed on the subset of workers S:

$$\mathbb{E}\left[T_{\text{time}}\right] = \mathcal{O}\left(\frac{\Delta}{\varepsilon} \max\left\{t(S)L_{\max}, d\kappa_{\max}(S)L_A\right\}\right).$$

Denote $\tilde{T}(S) := \max\left\{t(S)L_{\max}, d\kappa_{\max}(S)L_A\right\}$. Then, we define the optimal subset of workers as

$$S^* \in \operatorname*{argmin}_{\emptyset \neq S \subseteq [n]} \tilde{T}(S).$$

Note that S^* may not be unique. Theorem 4.3 provides guarantees for Algorithm 1.

Theorem 4.3. *A subset of workers S^* that minimizes*

$$\tilde{T}(S) := \max\left\{t(S)L_{\max}, d\max_{i \in S} \kappa_i L_A\right\}$$

over all $\emptyset \neq S \subseteq [n]$ can be found using Algorithm 1.

Proof. Assume the nonempty subset S minimizes $\tilde{T}(\cdot)$, i.e., there is no nonempty $S' \subseteq [n]$ such that $\tilde{T}(S') < \tilde{T}(S)$. Let π be a permutation of $\{1, \ldots, n\}$ such that $\kappa_{\pi(1)} \leq \kappa_{\pi(2)} \leq \cdots \leq \kappa_{\pi(n)}$. Define $i_S^\pi := \max\{i : \pi(i) \in S\}$. Consider the initial segment of sorted indices $\tilde{S} := \{\pi(1), \ldots, \pi(i_S^\pi)\}$; by construction, $S \subseteq \tilde{S}$.

Next, let σ be a permutation of the workers in $\tilde{S}$ such that $M_{\sigma(1)} \leq M_{\sigma(2)} \leq \cdots \leq M_{\sigma(i_S^\pi)}$. Define $i_S^\sigma := \max\{i : \sigma(i) \in S\}$ and consider the initial segment $\hat{S} := \{\sigma(1), \ldots, \sigma(i_S^\sigma)\}$. Again, $S \subseteq \hat{S}$ by construction.

Since the algorithm computes $\tilde{T}(\hat{S})/L_{\max}$ when $k = i_S^\pi$ and $m = i_S^\sigma$, the case $S = \hat{S}$ is trivial. Suppose instead that $S \subset \hat{S}$. Then there exists an index $j \leq i_S^\sigma$ such that $\sigma(j) \notin S$. By Lemma E.6, adding the worker $\sigma(j)$ to S does not increase the time complexity, i.e., $\tilde{T}(S \cup \{\sigma(j)\}) \leq \tilde{T}(S)$. By iteratively using Lemma E.6 and adding such workers from $\hat{S} \setminus S$, we can transform S into $\hat{S}$, which completes the proof. □

Lemma E.6 (Fast worker improves performance)**.** *Let $[n]$ be a set of workers' indices. Consider an arbitrary subset $\emptyset \neq S \subseteq [n]$ and $j \in [n]$ such that $j \notin S$, $M_j \leq M_{\max}(S)$, and $\kappa_j \leq \kappa_{\max}(S)$. Then, $\tilde{T}(S \cup \{j\}) \leq \tilde{T}(S)$.*

Proof. Since $M_j \leq M_{\max}(S)$ and $\kappa_j \leq \kappa_{\max}(S)$, we have

$$\begin{aligned}\tilde{T}(S \cup \{j\}) &= \max\left\{M_{\max}(S \cup \{j\})L_{\max}, s^*(S \cup \{j\})L_{\max}, d\kappa_{\max}(S \cup \{j\})L_A\right\} \\ &\leq \max\left\{M_{\max}(S)L_{\max}, s^*(S \cup \{j\})L_{\max}, d\kappa_{\max}(S)L_A\right\}.\end{aligned} \tag{33}$$

It remains to show $s^*(S \cup \{j\}) \leq s^*(S)$. Recall that $s^*(S)$ is the solution to the equation $\delta_S(s) = 1$. Comparing $\delta_S(s)$ and $\delta_{S \cup \{j\}}(s)$, we observe:

$$\begin{aligned}\delta_{S\cup\{j\}}(s) &= \left(\sum_{i \in S \cup \{j\}} \frac{1}{\frac{16\omega\tau_i}{s} + \frac{16\sigma^2 h_i}{\varepsilon s} + \frac{32\sigma^2\omega h_i\tau_i}{\varepsilon s^2} + \frac{4d\omega_{\text{s}}\kappa_{\max}(S\cup\{j\})\kappa_i}{s^2} + \frac{8d\omega_{\text{s}}\omega\kappa_{\max}(S\cup\{j\})\kappa_i\tau_i}{s^3}}\right)^{-1} \\ &= \left(\sum_{i \in S \cup \{j\}} \frac{1}{\frac{16\omega\tau_i}{s} + \frac{16\sigma^2 h_i}{\varepsilon s} + \frac{32\sigma^2\omega h_i\tau_i}{\varepsilon s^2} + \frac{4d\omega_{\text{s}}\kappa_{\max}(S)\kappa_i}{s^2} + \frac{8d\omega_{\text{s}}\omega\kappa_{\max}(S)\kappa_i\tau_i}{s^3}}\right)^{-1} \\ &= \left(\psi_S(s) + \frac{1}{\frac{16\omega\tau_j}{s} + \frac{16\sigma^2 h_j}{\varepsilon s} + \frac{32\sigma^2\omega h_j\tau_j}{\varepsilon s^2} + \frac{4d\omega_{\text{s}}\kappa_{\max}(S)\kappa_j}{s^2} + \frac{8d\omega_{\text{s}}\omega\kappa_{\max}(S)\kappa_j\tau_j}{s^3}}\right)^{-1} \leq \delta_S(s).\end{aligned}$$

We obtain that $\delta_{S\cup\{j\}}(s) \leq \delta_S(s)$ for all s. Thus, $\delta_{S\cup\{j\}}(s^*(S)) \leq \delta_S(s^*(S)) = 1$. Since $\delta_{S\cup\{j\}}(s)$ is a decreasing function and $\delta_{S\cup\{j\}}(s^*(S)) \leq \delta_{S\cup\{j\}}(s^*(S \cup \{j\})) = 1$, we conclude that $s^*(S \cup \{j\}) \leq s^*(S)$. Substituting the last inequality into (33) yields $\tilde{T}(S \cup \{j\}) \leq \tilde{T}(S)$. □

E.3 Bounding Variances: Auxiliary Lemmas

In this section we prove some inequalities to bound the residual between the gradient estimator g^k from (Heterogeneous-Time Inkheart SGD) and the true gradient and inequalities that control the distance between the point on the server x^k and local points x_i^k at which workers compute stochastic gradients.

Lemma E.7. *Consider the gradient estimator g^k defined in Heterogeneous-Time Inkheart SGD. Suppose that the function f satisfies Assumptions 1.1, 1.2, 1.3 and 1.6 and the compressors $\{\mathcal{C}_{ij}\} \in \mathbb{U}(\omega)$. Then, g^k is (in general) a biased estimator with $\mathbb{E}_k\left[g^k\right] = \sum_{i=1}^n \beta_i \nabla f(x_i^k)$ and the following inequality holds:*

$$\begin{aligned}
&\mathbb{E}_k\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \\
&\leq 4\left(\sum_{i=1}^n \beta_i^2 \frac{\omega}{m_i}\right)\left\|\nabla f(x^k)\right\|^2 + \sum_{i=1}^n \left(2L_A^2\beta_i + 4L^2\beta_i^2\frac{\omega}{m_i}\right)\left\|x_i^k - x^k\right\|^2 \\
&\quad + 2L_B^2\left\|\sum_{i=1}^n \beta_i\left(x_i^k - x^k\right)\right\|^2 + \sum_{i=1}^n 2\beta_i^2\left(\frac{\omega\sigma^2}{m_i b_i} + \frac{\sigma^2}{b_i}\right).
\end{aligned}$$

Proof. First, we compute the expectation. Since the compressors and the stochastic gradients are unbiased we obtain

$$\begin{aligned}
\mathbb{E}_k\left[g^k\right] &= \mathbb{E}_k\left[\sum_{i=1}^n \frac{\beta_i}{b_i m_i}\sum_{j=1}^{m_i}\mathcal{C}_{ij}\left(\sum_{r=1}^{b_i}\nabla f(x_i^k;\xi_{ir}^k)\right)\right] \\
&= \sum_{i=1}^n \frac{\beta_i}{b_i m_i}\sum_{j=1}^{m_i}\mathbb{E}_k\left[\mathbb{E}_{k,\xi}\left[\mathcal{C}_{ij}\left(\sum_{r=1}^{b_i}\nabla f(x_i^k;\xi_{ir}^k)\right)\right]\right] \\
&= \sum_{i=1}^n \frac{\beta_i}{b_i m_i}\sum_{j=1}^{m_i}\sum_{r=1}^{b_i}\mathbb{E}_k\left[\nabla f(x_i^k;\xi_{ir}^k)\right] = \sum_{i=1}^n \beta_i \nabla f(x_i^k).
\end{aligned}$$

Then, we consider the residual between the gradient estimator and the true gradient

$$\frac{1}{2}\mathbb{E}_k\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] = \frac{1}{2}\mathbb{E}_k\left[\left\|g^k - \sum_{i=1}^n \beta_i\nabla f(x_i^k) + \sum_{i=1}^n \beta_i \nabla f(x_i^k) - \nabla f(x^k)\right\|^2\right]$$

Then, we apply Young's inequality

$$\frac{1}{2}\mathbb{E}_k\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \leq \underbrace{\mathbb{E}_k\left[\left\|g^k - \sum_{i=1}^n \beta_i\nabla f(x_i^k)\right\|^2\right]}_{I_1} + \underbrace{\left\|\sum_{i=1}^n \beta_i\nabla f(x_i^k) - \nabla f(x^k)\right\|^2}_{I_2}.$$

Using Lemma E.8, we bound I_1:

$$I_1 \leq \sum_{i=1}^n \beta_i^2\left(2\frac{\omega}{m_i}\left\|\nabla f(x^k)\right\|^2 + 2L^2\frac{\omega}{m_i}\left\|x_i^k - x^k\right\|^2 + \frac{\omega\sigma^2}{m_i b_i} + \frac{\sigma^2}{b_i}\right). \tag{34}$$

To estimate I_2, we use Assumption C.2

$$I_2 \leq L_A^2\sum_{i=1}^n \beta_i\left\|x_i^k - x^k\right\|^2 + L_B^2\left\|\sum_{i=1}^n \beta_i\left(x_i^k - x^k\right)\right\|^2. \tag{35}$$

Combining (34) and (35), we derive

$$\frac{1}{2}\mathbb{E}_k\left[\left\|g^k - \nabla f(x^k)\right\|^2\right]$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{n} \beta_i^2 \left(2\frac{\omega}{m_i} \left\|\nabla f(x^k)\right\|^2 + 2L^2 \frac{\omega}{m_i} \left\|x_i^k - x^k\right\|^2 + \frac{\omega\sigma^2}{m_i b_i} + \frac{\sigma^2}{b_i} \right) \\
&\quad + L_A^2 \sum_{i=1}^{n} \beta_i \left\|x_i^k - x^k\right\|^2 + L_B^2 \left\| \sum_{i=1}^{n} \beta_i \left(x_i^k - x^k\right) \right\|^2 \\
&= \sum_{i=1}^{n} \beta_i^2 \left(2\frac{\omega}{m_i} \left\|\nabla f(x^k)\right\|^2 + \frac{\omega\sigma^2}{m_i b_i} + \frac{\sigma^2}{b_i} \right) \\
&\quad + \sum_{i=1}^{n} (L_A^2 \beta_i + 2L^2 \beta_i^2 \frac{\omega}{m_i}) \left\|x_i^k - x^k\right\|^2 + L_B^2 \left\| \sum_{i=1}^{n} \beta_i \left(x_i^k - x^k\right) \right\|^2 .
\end{aligned}
$$

□

The following lemma bounds the residual between the true gradients at the workers' local points and the gradient estimator.

Lemma E.8. *Consider the gradient estimator g^k defined in Heterogeneous-Time Inkheart SGD. Suppose that the function f satisfies Assumptions 1.1, 1.2, 1.3 and the compressors $\{\mathcal{C}_{ij}\} \in \mathbb{U}(\omega)$. Then, we have*

$$
\mathbb{E}_k \left[\left\| g^k - \sum_{i=1}^{n} \beta_i \nabla f(x_i^k) \right\|^2 \right] \leq \sum_{i=1}^{n} \beta_i^2 \left(2\frac{\omega}{m_i} \left\|\nabla f(x^k)\right\|^2 + 2L^2 \frac{\omega}{m_i} \left\|x_i^k - x^k\right\|^2 + \frac{\omega\sigma^2}{m_i b_i} + \frac{\sigma^2}{b_i} \right).
$$

Proof. Lemma E.7 yields $\mathbb{E}_k \left[g^k\right] = \sum_{i=1}^{n} \beta_i \nabla f(x_i^k)$. By the definition of the gradient estimator g^k in (Heterogeneous-Time Inkheart SGD), we obtain

$$
\begin{aligned}
\mathbb{E}_k \left[\left\| g^k - \sum_{i=1}^{n} \beta_i \nabla f(x_i^k) \right\|^2 \right] &= \mathbb{E}_k \left[\left\| \sum_{i=1}^{n} \frac{\beta_i}{b_i m_i} \sum_{j=1}^{m_i} \mathcal{C}_{ij} \left(\sum_{r=1}^{b_i} \nabla f(x_i^k; \xi_{ir}^k) \right) - \sum_{i=1}^{n} \beta_i \nabla f(x_i^k) \right\|^2 \right] \\
&= \sum_{i=1}^{n} \mathbb{E}_k \left[\left\| \frac{\beta_i}{b_i m_i} \sum_{j=1}^{m_i} \mathcal{C}_{ij} \left(\sum_{r=1}^{b_i} \nabla f(x_i^k; \xi_{ir}^k) \right) - \beta_i \nabla f(x_i^k) \right\|^2 \right]
\end{aligned}
$$

In the last equality we use the independence across workers. Next, we bound each term separately using the tower property and variance decomposition:

$$
\begin{aligned}
&\mathbb{E}_k \left[\left\| \frac{\beta_i}{b_i m_i} \sum_{j=1}^{m_i} \mathcal{C}_{ij} \left(\sum_{r=1}^{b_i} \nabla f(x_i^k; \xi_{ir}^k) \right) - \beta_i \nabla f(x_i^k) \right\|^2 \right] \\
&= \underbrace{\mathbb{E}_k \left[\left\| \frac{\beta_i}{b_i m_i} \sum_{j=1}^{m_i} \mathcal{C}_{ij} \left(\sum_{r=1}^{b_i} \nabla f(x_i^k; \xi_{ir}^k) \right) - \frac{\beta_i}{b_i m_i} \sum_{j=1}^{m_i} \sum_{r=1}^{b_i} \nabla f(x_i^k; \xi_{ir}^k) \right\|^2 \right]}_{I_1} \\
&\quad + \underbrace{\mathbb{E}_k \left[\left\| \frac{\beta_i}{b_i m_i} \sum_{j=1}^{m_i} \sum_{r=1}^{b_i} \nabla f(x_i^k; \xi_{ir}^k) - \beta_i \nabla f(x_i^k) \right\|^2 \right]}_{I_2}
\end{aligned}
$$

We analyze each term separately. We use the bounded variance property of compressors and variance decomposition.

$$
I_1 = \sum_{j=1}^{m_i} \left(\frac{\beta_i}{b_i m_i} \right)^2 \mathbb{E}_k \left[\left\| \mathcal{C}_{ij} \left(\sum_{r=1}^{b_i} \nabla f(x_i^k; \xi_{ir}^k) \right) - \sum_{r=1}^{b_i} \nabla f(x_i^k; \xi_{ir}^k) \right\|^2 \right]
$$

$$\leq \sum_{j=1}^{m_i} \left(\frac{\beta_i}{b_i m_i}\right)^2 \omega \mathbb{E}_k \left[\left\|\sum_{r=1}^{b_i} \nabla f(x_i^k; \xi_{ir}^k)\right\|^2\right] \leq \frac{\beta_i^2 \omega}{b_i^2 m_i} \left(b_i^2 \left\|\nabla f(x_i^k)\right\|^2 + b_i \sigma^2\right).$$

For each r, the stochastic gradients are independent. Therefore,

$$I_2 = \sum_{r=1}^{b_i} \mathbb{E}_k \left[\left\|\frac{\beta_i}{b_i m_i} \sum_{j=1}^{m_i} (\nabla f(x_i^k; \xi_{ir}^k) - \nabla f(x_i^k))\right\|^2\right]$$
$$= \sum_{r=1}^{b_i} \mathbb{E}_k \left[\left\|\frac{\beta_i}{b_i} (\nabla f(x_i^k; \xi_{ir}^k) - \nabla f(x_i^k))\right\|^2\right] \leq \frac{\beta_i^2 \sigma^2}{b_i}.$$

Thus, we obtain

$$\mathbb{E}_k \left[\left\|g^k - \sum_{i=1}^n \beta_i \nabla f(x_i^k)\right\|^2\right] \leq \sum_{i=1}^n \beta_i^2 \left(\frac{\omega}{m_i} \left\|\nabla f(x_i^k)\right\|^2 + \frac{\omega \sigma^2}{m_i b_i} + \frac{\sigma^2}{b_i}\right)$$
$$\leq \sum_{i=1}^n \beta_i^2 \left(2\frac{\omega}{m_i} \left\|\nabla f(x^k)\right\|^2 + 2\frac{\omega}{m_i} \left\|\nabla f(x_i^k) - \nabla f(x^k)\right\|^2 + \frac{\omega \sigma^2}{m_i b_i} + \frac{\sigma^2}{b_i}\right)$$
$$\leq \sum_{i=1}^n \beta_i^2 \left(2\frac{\omega}{m_i} \left\|\nabla f(x^k)\right\|^2 + 2L^2 \frac{\omega}{m_i} \left\|x_i^k - x^k\right\|^2 + \frac{\omega \sigma^2}{m_i b_i} + \frac{\sigma^2}{b_i}\right).$$

□

In the previous lemmas, we bounded the residual between the gradient estimate and the true gradient. In the following, we analyze the distance between the point at which the server performs the update and the local points at which the workers compute the stochastic gradients.

Lemma E.9. *Consider the point updates in Heterogeneous-Time Inkheart SGD. Suppose that the compressors $\mathcal{C}_{s,ij} \in \mathbb{U}(\omega_s)$, then the following inequalities hold:*

$$\sum_{i=1}^n \beta_i \mathbb{E}_k \left[\left\|x_i^{k+1} - x^{k+1}\right\|^2\right] \leq (1-p) \left[\sum_{i=1}^n \frac{\omega_s \beta_i}{\ell_i} \mathbb{E}_k \left[\left\|x^{k+1} - x^k\right\|^2\right] + \sum_{i=1}^n \beta_i \left\|x_i^k - x^k\right\|^2\right] \tag{36}$$

and

$$\mathbb{E}_k \left[\left\|\sum_{i=1}^n \beta_i \left(x_i^{k+1} - x^{k+1}\right)\right\|^2\right] \leq (1-p) \left[\sum_{i=1}^n \frac{\omega_s \beta_i^2}{\ell_i} \mathbb{E}_k \left[\left\|x^{k+1} - x^k\right\|^2\right] + \left\|\sum_{i=1}^n \beta_i \left(x_i^k - x^k\right)\right\|^2\right]. \tag{37}$$

Proof. We start by analyzing the evolution of the quantity $\left\|x_i^k - x^k\right\|^2$.

$$\mathbb{E}_k \left[\left\|x_i^{k+1} - x^{k+1}\right\|^2\right]$$
$$= \mathbb{E}_k \left[\mathbf{1}_{\{c^k=1\}} \left\|x^{k+1} - x^{k+1}\right\|^2 + \mathbf{1}_{\{c^k=0\}} \left\|x_i^k + \frac{1}{\ell_i} \sum_{j=1}^{\ell_i} \mathcal{C}_{s,ij}(x^{k+1} - x^k) - x^{k+1}\right\|^2\right]$$

Note that the first term vanishes. In the second term, c^k is independent of the other random variables and of the sigma-algebra associated with the conditional expectation $\mathbb{E}_k[\cdot]$. Thus, we can rewrite the expression as:

$$\mathbb{E}_k \left[\left\|x_i^{k+1} - x^{k+1}\right\|^2\right] = \mathbb{E}\left[\mathbf{1}_{\{c^k=0\}}\right] \cdot \mathbb{E}_k \left[\left\|x_i^k + \frac{1}{\ell_i} \sum_{j=1}^{\ell_i} \mathcal{C}_{s,ij}(x^{k+1} - x^k) - x^{k+1}\right\|^2\right]$$

$$= (1-p)\mathbb{E}_k\left[\mathbb{E}\left[\left\|\frac{1}{\ell_i}\sum_{j=1}^{\ell_i}\mathcal{C}_{\mathrm{s},ij}(x^{k+1}-x^k)-(x^{k+1}-x_i^k)\right\|^2\middle| x^{k+1}\right]\right]$$

Using Lemma D.2 for independent compressors we get

$$\begin{aligned}
&\mathbb{E}_k\left[\left\|x_i^{k+1}-x^{k+1}\right\|^2\right]\\
&= (1-p)\mathbb{E}_k\left[\mathbb{E}\left[\left\|\frac{1}{\ell_i}\sum_{j=1}^{\ell_i}\mathcal{C}_{\mathrm{s},ij}(x^{k+1}-x^k)-(x^{k+1}-x^k)\right\|^2\middle| x^{k+1}\right]\right]\\
&\quad+(1-p)\mathbb{E}_k\left[\left\|(x^{k+1}-x_i^k)-(x^{k+1}-x^k)\right\|^2\right]
\end{aligned}$$

Then, we apply Lemma D.4 to $\bar{\mathcal{C}}(x)=\frac{1}{\ell_i}\sum_{j=1}^{\ell_i}\mathcal{C}_{\mathrm{s},ij}(x)$

$$\mathbb{E}_k\left[\left\|x_i^{k+1}-x^{k+1}\right\|^2\right]\leq(1-p)\frac{\omega_{\mathrm{s}}}{\ell_i}\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right]+(1-p)\left\|x_i^k-x^k\right\|^2.$$

Proof for the second inequality is almost the same.

$$\mathbb{E}_k\left[\left\|\sum_{i=1}^n\beta_i\left(x_i^{k+1}-x^{k+1}\right)\right\|^2\right]=(1-p)\mathbb{E}_k\left[\left\|\sum_{i=1}^n\beta_i\left(\frac{1}{\ell_i}\sum_{j=1}^{\ell_i}\mathcal{C}_{\mathrm{s},ij}(x^{k+1}-x^k)-(x^{k+1}-x_i^k)\right)\right\|^2\right]$$

Using Lemma D.2 for independent compressors, we get

$$\begin{aligned}
&\mathbb{E}_k\left[\left\|\sum_{i=1}^n\beta_i\left(x_i^{k+1}-x^{k+1}\right)\right\|^2\right]\\
&=(1-p)\mathbb{E}_k\left[\mathbb{E}\left[\left\|\sum_{i=1}^n\beta_i\left(\frac{1}{\ell_i}\sum_{j=1}^{\ell_i}\mathcal{C}_{\mathrm{s},ij}(x^{k+1}-x^k)-(x^{k+1}-x^k)\right)\right\|^2\middle| x^{k+1}\right]\right]\\
&\quad+(1-p)\mathbb{E}_k\left[\left\|\sum_{i=1}^n\beta_i(x^{k+1}-x_i^k)-\sum_{i=1}^n\beta_i(x^{k+1}-x^k)\right\|^2\right]
\end{aligned}$$

Then, we apply Lemma D.4 to $\bar{\mathcal{C}}=\sum_{i=1}^n\sum_{j=1}^{\ell_i}\frac{\beta_i}{\ell_i}\mathcal{C}_{\mathrm{s},ij}$ and get

$$\begin{aligned}
&\mathbb{E}_k\left[\left\|\sum_{i=1}^n\beta_i\left(x_i^{k+1}-x^{k+1}\right)\right\|^2\right]\\
&\leq(1-p)\sum_{i=1}^n\sum_{j=1}^{\ell_i}\frac{\omega_{\mathrm{s}}\beta_i^2}{\ell_i^2}\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right]+(1-p)\mathbb{E}_k\left[\left\|\sum_{i=1}^n\beta_i(x_i^k-x^k)\right\|^2\right].
\end{aligned}$$

□

F Proofs for M4

We denote $v^k:=\frac{1}{n}\sum_{i=1}^n v_i^k$.

F.1 Known lemmas

We take several auxiliary lemmas from (Gruntkowska et al., 2024).

Lemma F.1 (Gruntkowska et al. (2024)). *Let $\{\mathcal{C}_{\mathrm{s},i}\}_{i=1}^n \in \mathbb{U}(\omega_{\mathrm{s}})$. Then, w_i^{k+1} in M4 satisfies*

$$\mathbb{E}_k\left[\left\|w_i^{k+1}-x_i^k\right\|^2\right] \leq \mathbb{E}_k\left[\left\|(x^{k+1}-x^k)-p_{\mathrm{s}}(w_i^k-x^k)+(w_i^k-x_i^k)\right\|^2\right] + \mathbb{E}_k\left[p_{\mathrm{s}}\left\|w_i^k-x^k\right\|^2\right]+\omega_{\mathrm{s}}\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right]$$

for all $i \in [n]$ and

$$\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n(w_i^{k+1}-x_i^k)\right\|^2\right] \leq \mathbb{E}_k\left[\left\|(x^{k+1}-x^k)-p_{\mathrm{s}}\left(\frac{1}{n}\sum_{i=1}^n w_i^k-x^k\right)+\frac{1}{n}\sum_{i=1}^n(w_i^k-x_i^k)\right\|^2\right] + p_{\mathrm{s}}\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n w_i^k-x^k\right\|^2\right]+\frac{\omega_{\mathrm{s}}}{n}\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right].$$

Lemma F.2 (Gruntkowska et al. (2024)). *Let $\{\mathcal{C}_{\mathrm{s},i}\}_{i=1}^n \in \mathbb{U}(\omega_{\mathrm{s}})$. Then, x_i^{k+1} in M4 satisfies*

$$\mathbb{E}_k\left[\left\|x_i^{k+1}-w_i^{k+1}\right\|^2\right] \leq \left(1-\frac{\mu}{2}\right)\mathbb{E}_k\left[\left\|w_i^k-x_i^k\right\|^2\right] + 4\left(\frac{1}{\mu}+\omega_{\mathrm{s}}\right)\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right] + 4p_{\mathrm{s}}\left(1+\frac{p_{\mathrm{s}}}{\mu}\right)\mathbb{E}_k\left[\left\|w_i^k-x^k\right\|^2\right]$$

and

$$\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n\left(x_i^{k+1}-w_i^{k+1}\right)\right\|^2\right] \leq \left(1-\frac{\mu}{2}\right)\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n\left(w_i^k-x_i^k\right)\right\|^2\right] + 4\left(\frac{1}{\mu}+\frac{\omega_{\mathrm{s}}}{n}\right)\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right] + 4p_{\mathrm{s}}\left(1+\frac{p_{\mathrm{s}}}{\mu}\right)\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n w_i^k-x^k\right\|^2\right]$$

Lemma F.3 (Gruntkowska et al. (2024)). *Let $\{\mathcal{C}_{\mathrm{s},i}\}_{i=1}^n \in \mathbb{U}(\omega_{\mathrm{s}})$. Then, w_i^{k+1} in M4 satisfies*

$$\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|w_i^{k+1}-x^{k+1}\right\|^2\right] \leq (1-p_{\mathrm{s}})\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|w_i^k-x^k\right\|^2\right]+(1-p_{\mathrm{s}})\omega_{\mathrm{s}}\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right]. \tag{38}$$

and

$$\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n w_i^{k+1}-x^{k+1}\right\|^2\right] \leq (1-p_{\mathrm{s}})\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n w_i^k-x^k\right\|^2\right]+(1-p_{\mathrm{s}})\frac{\omega_{\mathrm{s}}}{n}\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right]. \tag{39}$$

F.2 Auxiliary lemmas

Lemma F.4. *Consider M4. Let $\{\mathcal{C}_{\mathrm{s},i}\}_{i=1}^n \in \mathbb{U}(\omega_{\mathrm{s}})$. Then,*

$$\mathbb{E}_k\left[\left\|x_i^k-x_i^{k+1}\right\|^2\right] \leq 3\mu^2\mathbb{E}_k\left[\left\|w_i^k-x_i^k\right\|^2\right]+4\mu^2p_{\mathrm{s}}\mathbb{E}_k\left[\left\|w_i^k-x^k\right\|^2\right] + \mu^2(\omega_{\mathrm{s}}+3)\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right]$$

and

$$\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n\left(x_i^{k+1}-x_i^k\right)\right\|^2\right]\leq 3\mu^2\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n\left(w_i^k-x_i^k\right)\right\|^2\right]+4\mu^2p_{\rm s}\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n w_i^k-x^k\right\|^2\right]$$
$$+\mu^2\left(\frac{\omega_{\rm s}}{n}+3\right)\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right]$$

Proof. By the definition of x_i^{k+1} in M4, we obtain

$$\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n\left(x_i^{k+1}-x_i^k\right)\right\|^2\right]=\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n\left((1-\mu)x_i^k+\mu w_i^{k+1}-x_i^k\right)\right\|^2\right]$$
$$=\mu^2\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n\left(w_i^{k+1}-x_i^k\right)\right\|^2\right]$$

Then, we use Lemma F.1 and Jensen's inequality to obtain

$$\mathbb{E}_k\left[\left\|x_i^{k+1}-x_i^k\right\|^2\right]$$
$$\leq\mu^2\mathbb{E}_k\left[3\left\|w_i^k-x_i^k\right\|^2+3\left\|p_{\rm s}(w_i^k-x^k)\right\|^2+3\left\|x^{k+1}-x^k\right\|^2\right]$$
$$+\mu^2p_{\rm s}\mathbb{E}_k\left[\left\|w_i^k-x^k\right\|^2\right]+\mu^2\omega_{\rm s}\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right].$$

Using the inequality $p_{\rm s}^2\leq p_{\rm s}$ and simplifying we obtain the final bound. The proof of the second inequality is analogous. □

Lemma F.5. *Consider M4. Let $\{\mathcal{C}_i\}_{i=1}^n\in\mathbb{U}(\omega)$ and $\{\mathcal{C}_{{\rm s},i}\}_{i=1}^n\in\mathbb{U}(\omega_{\rm s})$. Let Assumptions 3.1, 3.2, and 3.3 be satisfied. Then,*

$$\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|v_i^{k+1}-v_i^k\right\|^2\right]\leq\frac{2\nu^2}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|v_i^k-\nabla f_i(x_i^k)\right\|^2\right]$$
$$+6\nu^2\mu^2L_{\max}^2\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|w_i^k-x_i^k\right\|^2\right]$$
$$+8\nu^2\mu^2p_{\rm s}L_{\max}^2\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|w_i^k-x^k\right\|^2\right]$$
$$+2\nu^2\mu^2(\omega_{\rm s}+3)\hat{L}^2\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right]+\nu^2\frac{\sigma^2}{b}$$

Proof. By the definition of v_i^{k+1} defined in M4, we get

$$\mathbb{E}_k\left[\left\|(1-\nu)v_i^k+\frac{\nu}{b}\sum_{r=1}^b\nabla f_i(x_i^{k+1};\xi_{ir}^{k+1})-v_i^k\right\|^2\right]$$
$$=\nu^2\mathbb{E}_k\left[\left\|\frac{1}{b}\sum_{r=1}^b\nabla f_i(x_i^{k+1};\xi_{ir}^{k+1})-v_i^k\right\|^2\right]$$
$$\overset{(i)}{=}\nu^2\mathbb{E}_k\left[\left\|\nabla f_i(x_i^{k+1})-\frac{1}{b}\sum_{r=1}^b\nabla f_i(x_i^{k+1};\xi_{ir}^{k+1})\right\|^2\right]+\nu^2\mathbb{E}_k\left[\left\|v_i^k-\nabla f_i(x_i^{k+1})\right\|^2\right]$$
$$\overset{(ii)}{\leq}\nu^2\frac{\sigma^2}{b}+\nu^2\mathbb{E}_k\left[\left\|v_i^k-\nabla f_i(x_i^{k+1})\right\|^2\right],$$

where (i) follows from the unbiasedness of the minibatch gradients conditional on the history and x_i^{k+1}, the tower property and the variance decomposition (Lemma D.2), while (ii) relies on Assumption 3.2. Consider the last term separately. Due to Young's inequality we obtain

$$\mathbb{E}_k\left[\left\|v_i^k - \nabla f_i(x_i^{k+1})\right\|^2\right] \leq 2\mathbb{E}_k\left[\left\|v_i^k - \nabla f_i(x_i^k)\right\|^2\right] + 2L_i^2\mathbb{E}_k\left[\left\|x_i^{k+1} - x_i^k\right\|^2\right]$$

Then, we apply Lemma F.4 to the last term and get

$$\begin{aligned}\mathbb{E}_k\left[\left\|v_i^{k+1} - v_i^k\right\|^2\right] &\leq 2\nu^2\mathbb{E}_k\left[\left\|v_i^k - \nabla f_i(x_i^k)\right\|^2\right] + 6\nu^2L_i^2\mu^2\mathbb{E}_k\left[\left\|w_i^k - x_i^k\right\|^2\right] \\ &\quad + 8\nu^2L_i^2\mu^2p_{\mathrm{s}}\mathbb{E}_k\left[\left\|w_i^k - x^k\right\|^2\right] + 2\nu^2L_i^2\mu^2(\omega_{\mathrm{s}}+3)\mathbb{E}_k\left[\left\|x^{k+1} - x^k\right\|^2\right] + \nu^2\frac{\sigma^2}{b}.\end{aligned}$$

Averaging over $i \in [n]$ and using

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n L_i^2\mathbb{E}_k\left[\left\|w_i^k - x_i^k\right\|^2\right] &\leq L_{\max}^2\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|w_i^k - x_i^k\right\|^2\right], \\ \frac{1}{n}\sum_{i=1}^n L_i^2\mathbb{E}_k\left[\left\|w_i^k - x^k\right\|^2\right] &\leq L_{\max}^2\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|w_i^k - x^k\right\|^2\right], \\ \text{and}\quad \frac{1}{n}\sum_{i=1}^n L_i^2\mathbb{E}_k\left[\left\|x^{k+1} - x^k\right\|^2\right] &= \hat{L}^2\mathbb{E}_k\left[\left\|x^{k+1} - x^k\right\|^2\right]\end{aligned}$$

we complete the proof. □

F.3 Bounding variances

First, we bound the deviation between the true gradient and the gradient estimator g^k from M4 in terms of simpler quantities.

Lemma F.6. *Consider M4. Then,*

$$\mathbb{E}_k\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \leq 3\mathbb{E}_k\left[\left\|g^k - v^k\right\|^2\right] + 3\mathbb{E}_k\left[\left\|v^k - \nabla f(x^k)\right\|^2\right].$$

Proof. Using simple algebra and Young's inequality we get

$$\begin{aligned}&\mathbb{E}_k\left[\left\|g^k - \nabla f(x^k)\right\|^2\right] \\ &= \mathbb{E}_k\left[\left\|g^k - v^k + v^k - \nabla f(x^k)\right\|^2\right] \\ &\leq 3\mathbb{E}_k\left[\left\|g^k - v^k\right\|^2\right] + 3\mathbb{E}_k\left[\left\|v^k - \nabla f(x^k)\right\|^2\right].\end{aligned}$$

□

The following lemma bounds the error induced by the worker compressors on average across workers.

Lemma F.7. *Consider M4. Let $\{\mathcal{C}_i\}_{i=1}^n \in \mathbb{U}(\omega)$, $\{\mathcal{C}_{\mathrm{s},i}\}_{i=1}^n \in \mathbb{U}(\omega_{\mathrm{s}})$. Let Assumptions 3.1, 3.2, and 3.3 be satisfied. Then,*

$$\begin{aligned}&\mathbb{E}_k\left[\left\|g^{k+1} - v^{k+1}\right\|^2\right] \\ &\leq (1-p)\mathbb{E}_k\left[\left\|g^k - v^k\right\|^2\right] \\ &\quad + \frac{2\nu^2\omega}{n}\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|v_i^k - \nabla f_i(x_i^k)\right\|^2\right] \\ &\quad + \frac{6\nu^2\mu^2\omega L_{\max}^2}{n}\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|w_i^k - x_i^k\right\|^2\right]\end{aligned}$$

$$+\frac{8\nu^2\mu^2 p_{\rm s}\omega L_{\max}^2}{n}\frac{1}{n}\sum_{i=1}^n \mathbb{E}_k\left[\left\|w_i^k - x^k\right\|^2\right]$$
$$+\frac{2\nu^2\mu^2(\omega_{\rm s}+3)\omega\hat{L}^2}{n}\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right]+\frac{\omega\nu^2\sigma^2}{nb}.$$

Proof. By the definition of g^{k+1} in M4, we get

$$\begin{aligned}
&\mathbb{E}_k\left[\left\|g^{k+1}-v^{k+1}\right\|^2\right]\\
&=(1-p)\mathbb{E}_k\left[\left\|g^k+\frac{1}{n}\sum_{i=1}^n \mathcal{C}_i(v_i^{k+1}-v_i^k)-v^{k+1}\right\|^2\right]\\
&\overset{(i)}{=}(1-p)\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n\left(\mathcal{C}_i(v_i^{k+1}-v_i^k)-(v_i^{k+1}-v_i^k)\right)\right\|^2\right]+(1-p)\mathbb{E}_k\left[\left\|g^k-v^k\right\|^2\right]\\
&\overset{(ii)}{=}(1-p)\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}_k\left[\left\|\mathcal{C}_i(v_i^{k+1}-v_i^k)-\left(v_i^{k+1}-v_i^k\right)\right\|^2\right]+(1-p)\mathbb{E}_k\left[\left\|g^k-v^k\right\|^2\right]\\
&\le(1-p)\mathbb{E}_k\left[\left\|g^k-v^k\right\|^2\right]+\frac{(1-p)\omega}{n}\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|v_i^{k+1}-v_i^k\right\|^2\right],
\end{aligned}$$

where (i) follows from the unbiasedness of the compressors, the tower property and the variance decomposition (Lemma D.2), while (ii) relies on the fact that the compressors are independent. Then, we use Lemma F.5 and obtain

$$\begin{aligned}
&\mathbb{E}_k\left[\left\|g^{k+1}-v^{k+1}\right\|^2\right]\\
&\le(1-p)\mathbb{E}_k\left[\left\|g^k-v^k\right\|^2\right]\\
&\quad+(1-p)\frac{2\nu^2\omega}{n}\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|v_i^k-\nabla f_i(x_i^k)\right\|^2\right]\\
&\quad+(1-p)\frac{6\nu^2\mu^2\omega L_{\max}^2}{n}\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|w_i^k-x_i^k\right\|^2\right]\\
&\quad+(1-p)\frac{8\nu^2\mu^2 p_{\rm s}\omega L_{\max}^2}{n}\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|w_i^k-x^k\right\|^2\right]\\
&\quad+(1-p)\frac{2\nu^2\mu^2(\omega_{\rm s}+3)\omega\hat{L}^2}{n}\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right]+\frac{(1-p)\omega\nu^2\sigma^2}{nb}.
\end{aligned}$$

Applying the inequality $1-p\le 1$ we complete the proof. □

The following lemma controls the bias caused by the momentum defined in M4.

Lemma F.8. *Let $\{\mathcal{C}_i\}_{i=1}^n\in\mathbb{U}(\omega)$, $\{\mathcal{C}_{{\rm s},i}\}_{i=1}^n\in\mathbb{U}(\omega_{\rm s})$. Let Assumptions 3.1, 3.2, and 3.3 be satisfied. Then,*

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|v_i^{k+1}-\nabla f_i(x_i^{k+1})\right\|^2\right]&\le(1-\nu)\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|v_i^k-\nabla f_i(x_i^k)\right\|^2\right]\\
&\quad+\frac{9L_{\max}^2\mu^2}{\nu}\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|w_i^k-x_i^k\right\|^2\right]\\
&\quad+\frac{12L_{\max}^2\mu^2 p_{\rm s}}{\nu}\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|w_i^k-x^k\right\|^2\right]
\end{aligned}$$

$$+\frac{3\hat{L}^2\mu^2(\omega_{\rm s}+3)}{\nu}\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right]+\nu^2\frac{\sigma^2}{b}$$

and

$$\begin{aligned}
&\mathbb{E}_k\left[\left\|v^{k+1}-\nabla f(x^{k+1})\right\|^2\right]\\
&\le(1-\nu)\mathbb{E}_k\left[\left\|v^k-\nabla f(x^k)\right\|^2\right]\\
&\quad+4\nu L_A^2\left(1-\frac{\mu}{2}\right)\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|w_i^k-x_i^k\right\|^2\right]+16\nu L_A^2\left(1+\frac{p_{\rm s}}{\mu}\right)\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|w_i^k-x^k\right\|^2\right]\\
&\quad+4\nu L_B^2\left(1-\frac{\mu}{2}\right)\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n(w_i^k-x_i^k)\right\|^2\right]+16\nu L_B^2\left(1+\frac{p_{\rm s}}{\mu}\right)\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n w_i^k-x^k\right\|^2\right]\\
&\quad+\left(\frac{2L^2}{\nu}+4\nu\left(\frac{4}{\mu}+5\omega_{\rm s}\right)L_A^2+4\nu\left(\frac{4}{\mu}+\frac{5\omega_{\rm s}}{n}\right)L_B^2\right)\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right]+\frac{\nu^2\sigma^2}{nb}.
\end{aligned}$$

Proof. We start with the first inequality. By the definition of v_i^{k+1} in (M4) we get

$$\begin{aligned}
&\mathbb{E}_k\left[\left\|v_i^{k+1}-\nabla f_i(x_i^{k+1})\right\|^2\right]\\
&=\mathbb{E}_k\left[\left\|(1-\nu)v_i^k+\frac{\nu}{b}\sum_{r=1}^b\nabla f_i(x_i^{k+1};\xi_{ir}^{k+1})-\nabla f_i(x_i^{k+1})\right\|^2\right]\\
&=\mathbb{E}_k\left[\left\|(1-\nu)\left(v_i^k-\nabla f_i(x_i^{k+1})\right)+\nu\left(\frac{1}{b}\sum_{r=1}^b\nabla f_i(x_i^{k+1};\xi_{ir}^{k+1})-\nabla f_i(x_i^{k+1})\right)\right\|^2\right]\\
&\overset{(i)}{=}(1-\nu)^2\mathbb{E}_k\left[\left\|v_i^k-\nabla f_i(x_i^{k+1})\right\|^2\right]+\nu^2\mathbb{E}_k\left[\left\|\frac{1}{b}\sum_{r=1}^b\nabla f_i(x_i^{k+1};\xi_{ir}^{k+1})-\nabla f_i(x_i^{k+1})\right\|^2\right]\\
&\overset{(ii)}{\le}(1-\nu)^2(1+\rho)\mathbb{E}_k\left[\left\|v_i^k-\nabla f_i(x_i^k)\right\|^2\right]\\
&\quad+(1-\nu)^2\left(1+\frac{1}{\rho}\right)\mathbb{E}_k\left[\left\|\nabla f_i(x_i^k)-\nabla f_i(x_i^{k+1})\right\|^2\right]+\nu^2\frac{\sigma^2}{b}\\
&\le(1-\nu)^2(1+\rho)\mathbb{E}_k\left[\left\|v_i^k-\nabla f_i(x_i^k)\right\|^2\right]+(1-\nu)^2\left(1+\frac{1}{\rho}\right)L_i^2\mathbb{E}_k\left[\left\|x_i^k-x_i^{k+1}\right\|^2\right]+\nu^2\frac{\sigma^2}{b}\\
&\overset{(iii)}{\le}(1-\nu)\mathbb{E}_k\left[\left\|v_i^k-\nabla f_i(x_i^k)\right\|^2\right]+\frac{3L_i^2}{\nu}\mathbb{E}_k\left[\left\|x_i^k-x_i^{k+1}\right\|^2\right]+\nu^2\frac{\sigma^2}{b},
\end{aligned}$$

where (i) follows from the variance decomposition (Lemma D.2) and Assumption 3.2. (ii) applies Young's inequality with parameter $\rho>0$. (iii) is obtained by setting $\rho=\frac{\nu}{2}$. Then, use Lemma F.4

$$\begin{aligned}
&\mathbb{E}_k\left[\left\|v_i^{k+1}-\nabla f_i(x_i^{k+1})\right\|^2\right]\\
&\le(1-\nu)\mathbb{E}_k\left[\left\|v_i^k-\nabla f_i(x_i^k)\right\|^2\right]+\frac{9L_i^2\mu^2}{\nu}\mathbb{E}_k\left[\left\|w_i^k-x_i^k\right\|^2\right]+\frac{12L_i^2\mu^2p_{\rm s}}{\nu}\mathbb{E}_k\left[\left\|w_i^k-x^k\right\|^2\right]\\
&\quad+\frac{3L_i^2\mu^2(\omega_{\rm s}+3)}{\nu}\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right]+\nu^2\frac{\sigma^2}{b}.
\end{aligned}$$

Averaging and using

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n L_i^2\mathbb{E}_k\left[\left\|w_i^k-x_i^k\right\|^2\right]&\le L_{\max}^2\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|w_i^k-x_i^k\right\|^2\right],\\
\frac{1}{n}\sum_{i=1}^n L_i^2\mathbb{E}_k\left[\left\|w_i^k-x^k\right\|^2\right]&\le L_{\max}^2\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\left[\left\|w_i^k-x^k\right\|^2\right],
\end{aligned}$$

$$\text{and} \quad \frac{1}{n}\sum_{i=1}^{n} L_i^2 \mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right] = \hat{L}^2 \mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right],$$

we complete the proof.

Then, we consider the second inequality.

$$\begin{aligned}
&\mathbb{E}_k\left[\left\|v^{k+1}-\nabla f(x^{k+1})\right\|^2\right] \\
&= \mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\left((1-\nu)v_i^k + \frac{\nu}{b}\sum_{r=1}^{b}\nabla f_i(x_i^{k+1};\xi_{ir}^{k+1})\right) - \nabla f(x^{k+1})\right\|^2\right] \\
&\overset{(i)}{=} \mathbb{E}_k\left[\left\|(1-\nu)v^k + \nu\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i^{k+1}) - \nabla f(x^{k+1})\right\|^2\right] \\
&\quad + \nu^2\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{b}\sum_{r=1}^{b}\nabla f_i(x_i^{k+1};\xi_{ir}^{k+1}) - \nabla f_i(x_i^{k+1})\right)\right\|^2\right] \\
&\overset{(ii)}{=} \mathbb{E}_k\left[\left\|(1-\nu)\left(v^k-\nabla f(x^k)\right) + (1-\nu)\left(\nabla f(x^k)-\nabla f(x^{k+1})\right) + \nu\left(\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i^{k+1}) - \nabla f(x^{k+1})\right)\right\|^2\right] \\
&\quad + \frac{\nu^2}{n^2}\sum_{i=1}^{n}\mathbb{E}_k\left[\left\|\frac{1}{b}\sum_{r=1}^{b}\nabla f_i(x_i^{k+1};\xi_{ir}^{k+1}) - \nabla f_i(x_i^{k+1})\right\|^2\right] \\
&\overset{(iii)}{\leq} (1-\nu)\mathbb{E}_k\left[\left\|v^k-\nabla f(x^k)\right\|^2\right] \\
&\quad + \nu\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i^{k+1}) - \nabla f(x^{k+1}) + \frac{1-\nu}{\nu}\left(\nabla f(x^k)-\nabla f(x^{k+1})\right)\right\|^2\right] + \frac{\nu^2\sigma^2}{nb} \\
&\overset{(iv)}{\leq} (1-\nu)\mathbb{E}_k\left[\left\|v^k-\nabla f(x^k)\right\|^2\right] + \frac{2(1-\nu)^2}{\nu}\mathbb{E}_k\left[\left\|\nabla f(x^k)-\nabla f(x^{k+1})\right\|^2\right] \\
&\quad + 2\nu\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i^{k+1}) - \nabla f(x^{k+1})\right\|^2\right] + \frac{\nu^2\sigma^2}{nb} \\
&\leq (1-\nu)\mathbb{E}_k\left[\left\|v^k-\nabla f(x^k)\right\|^2\right] + \frac{2L^2}{\nu}\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right] \\
&\quad + 2\nu\underbrace{\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i^{k+1}) - \nabla f(x^{k+1})\right\|^2\right]}_{I} + \frac{\nu^2\sigma^2}{nb},
\end{aligned}$$

where (i) follows from the variance decomposition (Lemma D.2) and the unbiasedness of the minibatch gradients conditional on the history and the new local points. (ii) follows from Lemma D.3. (iii) uses convexity of the squared norm with weights $1-\nu$ and ν, and Assumption 3.2. (iv) applies Young's inequality. The last inequality uses Assumption 3.1 and $(1-\nu)^2 \leq 1$. We consider I separately. Using Assumption 3.3 with the common point x^{k+1}, we get

$$\begin{aligned}
I &= \mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla f_i(x_i^{k+1}) - \nabla f_i(x^{k+1})\right)\right\|^2\right] \\
&\leq L_A^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_k\left[\left\|x_i^{k+1}-x^{k+1}\right\|^2\right] + L_B^2\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^{n}x_i^{k+1} - x^{k+1}\right\|^2\right]
\end{aligned}$$

$$\leq 2L_A^2\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_k\left[\left\|w_i^{k+1}-x_i^{k+1}\right\|^2\right]+\frac{1}{n}\sum_{i=1}^n \mathbb{E}_k\left[\left\|w_i^{k+1}-x^{k+1}\right\|^2\right]\right)$$
$$+2L_B^2\left(\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n(w_i^{k+1}-x_i^{k+1})\right\|^2\right]+\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n w_i^{k+1}-x^{k+1}\right\|^2\right]\right).$$

Here we use Young's inequality twice. Then, applying Lemmas F.2 and F.3, we obtain

$$\begin{aligned}
I &\leq 2L_A^2\left(1-\frac{\mu}{2}\right)\frac{1}{n}\sum_{i=1}^n \mathbb{E}_k\left[\left\|w_i^k-x_i^k\right\|^2\right]+2L_A^2\left(4p_{\mathrm{s}}\left(1+\frac{p_{\mathrm{s}}}{\mu}\right)+1-p_{\mathrm{s}}\right)\frac{1}{n}\sum_{i=1}^n \mathbb{E}_k\left[\left\|w_i^k-x^k\right\|^2\right]\\
&\quad+2L_B^2\left(1-\frac{\mu}{2}\right)\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n(w_i^k-x_i^k)\right\|^2\right]+2L_B^2\left(4p_{\mathrm{s}}\left(1+\frac{p_{\mathrm{s}}}{\mu}\right)+1-p_{\mathrm{s}}\right)\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n w_i^k-x^k\right\|^2\right]\\
&\quad+2L_A^2\left(4\left(\frac{1}{\mu}+\omega_{\mathrm{s}}\right)+(1-p_{\mathrm{s}})\omega_{\mathrm{s}}\right)\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right]\\
&\quad+2L_B^2\left(4\left(\frac{1}{\mu}+\frac{\omega_{\mathrm{s}}}{n}\right)+(1-p_{\mathrm{s}})\frac{\omega_{\mathrm{s}}}{n}\right)\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right]\\
&\leq 2L_A^2\left(1-\frac{\mu}{2}\right)\frac{1}{n}\sum_{i=1}^n \mathbb{E}_k\left[\left\|w_i^k-x_i^k\right\|^2\right]+8L_A^2\left(1+\frac{p_{\mathrm{s}}}{\mu}\right)\frac{1}{n}\sum_{i=1}^n \mathbb{E}_k\left[\left\|w_i^k-x^k\right\|^2\right]\\
&\quad+2L_B^2\left(1-\frac{\mu}{2}\right)\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n(w_i^k-x_i^k)\right\|^2\right]+8L_B^2\left(1+\frac{p_{\mathrm{s}}}{\mu}\right)\mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n w_i^k-x^k\right\|^2\right]\\
&\quad+\left(2L_A^2\left(\frac{4}{\mu}+5\omega_{\mathrm{s}}\right)+2L_B^2\left(\frac{4}{\mu}+\frac{5\omega_{\mathrm{s}}}{n}\right)\right)\mathbb{E}_k\left[\left\|x^{k+1}-x^k\right\|^2\right].
\end{aligned}$$

Here we use $1+3p_{\mathrm{s}}+\frac{4p_{\mathrm{s}}^2}{\mu}\leq 4\left(1+\frac{p_{\mathrm{s}}}{\mu}\right)$ and $1-p_{\mathrm{s}}\leq 1$, since $p_{\mathrm{s}}\leq 1$. Substituting this bound into the preceding inequality completes the proof. □

Theorem F.9 (General theorem). *Assume that the function f satisfies Assumptions 3.1, 1.2, 3.2 and 3.3. Suppose that $\{\mathcal{C}_i\}_{i=1}^n\in\mathbb{U}(\omega)$ and $\{\mathcal{C}_{\mathrm{s},i}\}_{i=1}^n\in\mathbb{U}(\omega_{\mathrm{s}})$. Let $\gamma>0$ be such that*

$$\gamma\leq\left(L+\sqrt{c\left(\tilde{C}_A L_A^2+\tilde{C}_B L_B^2+\tilde{C}_{\max}L_{\max}^2\right)}\right)^{-1}, \tag{40}$$

where $c=1416$,

$$\begin{aligned}
\tilde{C}_A&=\left(\frac{\omega_{\mathrm{s}}p_{\mathrm{s}}+1}{\mu^2}+\frac{\omega_{\mathrm{s}}}{p_{\mathrm{s}}}\right),\\
\tilde{C}_B&=\left(\frac{\frac{\omega_{\mathrm{s}}}{n}p_{\mathrm{s}}+1}{\mu^2}+\frac{\omega_{\mathrm{s}}}{np_{\mathrm{s}}}\right),\\
\text{and}\quad \tilde{C}_{\max}&=\left(\frac{\mu\omega\omega_{\mathrm{s}}}{np}+\frac{\omega(1+\omega_{\mathrm{s}}p_{\mathrm{s}})}{np}+\frac{1}{\nu^2}\right).
\end{aligned}$$

Let

$$\begin{aligned}
\Psi^k&=f(x^k)-f^*+\lambda_A\left\|g^k-v^k\right\|^2+\lambda_B\left\|v^k-\nabla f(x^k)\right\|^2+\lambda_C\frac{1}{n}\sum_{i=1}^n\left\|v_i^k-\nabla f_i(x_i^k)\right\|^2\\
&\quad+\lambda_D\frac{1}{n}\sum_{i=1}^n\left\|w_i^k-x_i^k\right\|^2+\lambda_E\frac{1}{n}\sum_{i=1}^n\left\|w_i^k-x^k\right\|^2\\
&\quad+\lambda_F\left\|\frac{1}{n}\sum_{i=1}^n\left(w_i^k-x_i^k\right)\right\|^2+\lambda_G\left\|\frac{1}{n}\sum_{i=1}^n w_i^k-x^k\right\|^2,
\end{aligned}$$

where $\lambda_A = \frac{3\gamma}{2p}$, $\lambda_B = \frac{3\gamma}{2\nu}$, $\lambda_C = \frac{3\nu\omega\gamma}{np}$, $\lambda_D = \gamma\left(\frac{72\mu L_{\max}^2\omega}{np} + \frac{12L_A^2}{\mu}\right)$, $\lambda_E = \gamma\left(\frac{336\omega L_{\max}^2}{np}(\mu + p_{\rm s}) + 96L_A^2\left(\frac{1}{p_{\rm s}} + \frac{p_{\rm s}}{\mu^2}\right)\right)$, $\lambda_F = \frac{12\gamma L_B^2}{\mu}$ *and* $\lambda_G = 96\gamma L_B^2\left(\frac{p_{\rm s}}{\mu^2} + \frac{1}{p_{\rm s}}\right)$. *Then, M4 with* ν *such that*

$$\frac{9\nu^2\omega\sigma^2}{npb} + \frac{3\nu\sigma^2}{nb} \leq \frac{\varepsilon}{2}$$

ensures that

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(x^k)\right\|^2\right] \leq \frac{2\mathbb{E}\left[\Psi^0\right]}{\gamma K} + \frac{\varepsilon}{2}.$$

Proof. Using Assumption 3.1 and Lemma D.9,

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

Using Lemma F.6 and taking expectation $\mathbb{E}_k[\cdot]$ from both parts, we obtain

$$\begin{aligned}\mathbb{E}_k\left[f(x^{k+1})\right] \leq f(x^k) - \frac{\gamma}{2}\left\|\nabla f(x^k)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\mathbb{E}_k\left[\left\|x^{k+1} - x^k\right\|^2\right] \\ + \frac{3\gamma}{2}\mathbb{E}_k\left[\left\|g^k - v^k\right\|^2\right] + \frac{3\gamma}{2}\mathbb{E}_k\left[\left\|v^k - \nabla f(x^k)\right\|^2\right].\end{aligned} \tag{41}$$

To simplify the derivation, we define the seven norm-based terms in the Lyapunov function Ψ^k as follows:

$$\begin{aligned}
&\delta^k = f(x^k) - f^* && X^k = \mathbb{E}_k\left[\left\|x^{k+1} - x^k\right\|^2\right] \\
&A^k = \mathbb{E}_k\left[\left\|g^k - v^k\right\|^2\right] && B^k = \mathbb{E}_k\left[\left\|v^k - \nabla f(x^k)\right\|^2\right] \\
&C^k = \mathbb{E}_k\left[\frac{1}{n}\sum_{i=1}^n\left\|v_i^k - \nabla f_i(x_i^k)\right\|^2\right] && D^k = \mathbb{E}_k\left[\frac{1}{n}\sum_{i=1}^n\left\|w_i^k - x_i^k\right\|^2\right] \\
&E^k = \mathbb{E}_k\left[\frac{1}{n}\sum_{i=1}^n\left\|w_i^k - x^k\right\|^2\right] && F^k = \mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n\left(w_i^k - x_i^k\right)\right\|^2\right] \\
&G^k = \mathbb{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n w_i^k - x^k\right\|^2\right]
\end{aligned}$$

We define $\tilde{A}^{k+1} = \mathbb{E}_k\left[A^{k+1}\right], \ldots, \tilde{G}^{k+1} = \mathbb{E}_k\left[G^{k+1}\right]$. Using the notation introduced above, inequality (41) takes the form

$$\mathbb{E}_k\left[\delta^{k+1}\right] + \frac{\gamma}{2}\left\|\nabla f(x^k)\right\|^2 \leq \delta^k + \frac{3\gamma}{2}A^k + \frac{3\gamma}{2}B^k - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)X^k. \tag{42}$$

Adding $\mathbb{E}_k\left[\Psi^{k+1} - \delta^{k+1}\right]$ to both sides of inequality (42) yields

$$\mathbb{E}_k\left[\Psi^{k+1}\right] + \frac{\gamma}{2}\left\|\nabla f(x^k)\right\|^2 \leq \mathbb{E}_k\left[\Psi^{k+1} - \delta^{k+1}\right] + \delta^k + \frac{3\gamma}{2}A^k + \frac{3\gamma}{2}B^k - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)X^k. \tag{43}$$

In this notation, the recursion lemmas take the following form:

1. Lemma (F.7) yields

$$\begin{aligned}\tilde{A}^{k+1} \leq (1-p)A^k + \frac{2\nu^2\omega}{n}C^k + \frac{6\nu^2\mu^2\omega L_{\max}^2}{n}D^k + \frac{8\nu^2\mu^2 p_{\rm s}\omega L_{\max}^2}{n}E^k \\ + \frac{2\nu^2\mu^2(\omega_{\rm s}+3)\omega\hat{L}^2}{n}X^k + \frac{\omega\nu^2\sigma^2}{nb}.\end{aligned} \tag{44}$$

2. Lemma F.8 implies

$$\tilde{C}^{k+1} \leq (1-\nu)C^k + \frac{9L_{\max}^2\mu^2}{\nu}D^k + \frac{12L_{\max}^2\mu^2 p_{\rm s}}{\nu}E^k + \frac{3\hat{L}^2\mu^2(\omega_{\rm s}+3)}{\nu}X^k + \nu^2\frac{\sigma^2}{b}. \tag{45}$$

3. Lemma F.8 yields

$$\begin{aligned}\tilde{B}^{k+1} \leq{}& (1-\nu)B^k + 4\nu L_A^2\left(1-\frac{\mu}{2}\right)D^k + 16\nu L_A^2\left(1+\frac{p_{\rm s}}{\mu}\right)E^k \\ &+ 4\nu L_B^2\left(1-\frac{\mu}{2}\right)F^k + 16\nu L_B^2\left(1+\frac{p_{\rm s}}{\mu}\right)G^k \\ &+ \left(\frac{2L^2}{\nu} + 4\nu\left(\frac{4}{\mu}+5\omega_{\rm s}\right)L_A^2 + 4\nu\left(\frac{4}{\mu}+\frac{5\omega_{\rm s}}{n}\right)L_B^2\right)X^k + \frac{\nu^2\sigma^2}{nb}.\end{aligned} \tag{46}$$

4. Lemma F.2 gives

$$\tilde{D}^{k+1} \leq \left(1-\frac{\mu}{2}\right)D^k + 4p_{\rm s}\left(1+\frac{p_{\rm s}}{\mu}\right)E^k + 4\left(\frac{1}{\mu}+\omega_{\rm s}\right)X^k. \tag{47}$$

5. Lemma F.2 implies

$$\tilde{F}^{k+1} \leq \left(1-\frac{\mu}{2}\right)F^k + 4p_{\rm s}\left(1+\frac{p_{\rm s}}{\mu}\right)G^k + 4\left(\frac{1}{\mu}+\frac{\omega_{\rm s}}{n}\right)X^k. \tag{48}$$

6. Lemma F.3 where we additionally use the inequality $1-p_{\rm s} \leq 1$ gives

$$\tilde{E}^{k+1} \leq (1-p_{\rm s})E^k + \omega_{\rm s}X^k. \tag{49}$$

7. Lemma F.3 where we additionally use the inequality $1-p_{\rm s} \leq 1$ yields

$$\tilde{G}^{k+1} \leq (1-p_{\rm s})G^k + \frac{\omega_{\rm s}}{n}X^k. \tag{50}$$

Substituting inequalities (44)–(50) into the definition of Ψ^{k+1}, we obtain:

$$\begin{aligned}&\mathbb{E}_k\left[\Psi^{k+1} - \delta^{k+1}\right] \\ &= \lambda_A\tilde{A}^{k+1} + \lambda_B\tilde{B}^{k+1} + \lambda_C\tilde{C}^{k+1} + \lambda_D\tilde{D}^{k+1} + \lambda_E\tilde{E}^{k+1} + \lambda_F\tilde{F}^{k+1} + \lambda_G\tilde{G}^{k+1} \\ &\leq \lambda_A\left[(1-p)A^k + \frac{2\nu^2\omega}{n}C^k + \frac{6\nu^2\mu^2\omega L_{\max}^2}{n}D^k + \frac{8\nu^2\mu^2 p_{\rm s}\omega L_{\max}^2}{n}E^k + \frac{2\nu^2\mu^2(\omega_{\rm s}+3)\omega\hat{L}^2}{n}X^k + \frac{\omega\nu^2\sigma^2}{nb}\right] \\ &\quad + \lambda_B\left[(1-\nu)B^k + 4\nu L_A^2\left(1-\frac{\mu}{2}\right)D^k + 16\nu L_A^2\left(1+\frac{p_{\rm s}}{\mu}\right)E^k\right] \\ &\quad + \lambda_B\left[4\nu L_B^2\left(1-\frac{\mu}{2}\right)F^k + 16\nu L_B^2\left(1+\frac{p_{\rm s}}{\mu}\right)G^k\right] \\ &\quad + \lambda_B\left[\left(\frac{2L^2}{\nu} + 4\nu\left(\frac{4}{\mu}+5\omega_{\rm s}\right)L_A^2 + 4\nu\left(\frac{4}{\mu}+\frac{5\omega_{\rm s}}{n}\right)L_B^2\right)X^k + \frac{\nu^2\sigma^2}{nb}\right] \\ &\quad + \lambda_C\left[(1-\nu)C^k + \frac{9L_{\max}^2\mu^2}{\nu}D^k + \frac{12L_{\max}^2\mu^2 p_{\rm s}}{\nu}E^k + \frac{3\hat{L}^2\mu^2(\omega_{\rm s}+3)}{\nu}X^k + \frac{\nu^2\sigma^2}{b}\right] \\ &\quad + \lambda_D\left[\left(1-\frac{\mu}{2}\right)D^k + 4p_{\rm s}\left(1+\frac{p_{\rm s}}{\mu}\right)E^k + 4\left(\frac{1}{\mu}+\omega_{\rm s}\right)X^k\right] \\ &\quad + \lambda_E\left[(1-p_{\rm s})E^k + \omega_{\rm s}X^k\right] \\ &\quad + \lambda_F\left[\left(1-\frac{\mu}{2}\right)F^k + 4p_{\rm s}\left(1+\frac{p_{\rm s}}{\mu}\right)G^k + 4\left(\frac{1}{\mu}+\frac{\omega_{\rm s}}{n}\right)X^k\right]\end{aligned}$$

$$+\lambda_G\left[(1-p_{\rm s})G^k+\frac{\omega_{\rm s}}{n}X^k\right].$$

Next, we group the terms corresponding to the quantities $(A^k, B^k, \ldots, G^k, X^k)$ and the constant terms multiplied by σ^2. The resulting inequality can be written as:

$$\begin{aligned}&\mathbb{E}_k\left[\Psi^{k+1}-\delta^{k+1}\right]\\&\leq\tilde{\lambda}_A A^k+\tilde{\lambda}_B B^k+\tilde{\lambda}_C C^k+\tilde{\lambda}_D D^k+\tilde{\lambda}_E E^k+\tilde{\lambda}_F F^k+\tilde{\lambda}_G G^k+\tilde{\lambda}_X X^k+\tilde{c},\end{aligned}\tag{51}$$

where the coefficients $\tilde{\lambda}_A, \ldots, \tilde{\lambda}_G, \tilde{\lambda}_X, \tilde{c}$ are defined as follows:

$$\tilde{\lambda}_A=\lambda_A(1-p),\tag{52}$$

$$\tilde{\lambda}_B=\lambda_B(1-\nu),\tag{53}$$

$$\tilde{\lambda}_C=\lambda_C(1-\nu)+\lambda_A\frac{2\nu^2\omega}{n},\tag{54}$$

$$\tilde{\lambda}_D=\lambda_D\left(1-\frac{\mu}{2}\right)+\lambda_A\frac{6\nu^2\mu^2\omega L_{\max}^2}{n}+\lambda_B 4\nu L_A^2\left(1-\frac{\mu}{2}\right)+\lambda_C\frac{9L_{\max}^2\mu^2}{\nu},\tag{55}$$

$$\begin{aligned}\tilde{\lambda}_E=&\lambda_E(1-p_{\rm s})+\lambda_A\frac{8\nu^2\mu^2 p_{\rm s}\omega L_{\max}^2}{n}+\lambda_B 16\nu L_A^2\left(1+\frac{p_{\rm s}}{\mu}\right)\\&+\lambda_C\frac{12L_{\max}^2\mu^2 p_{\rm s}}{\nu}+\lambda_D 4p_{\rm s}\left(1+\frac{p_{\rm s}}{\mu}\right),\end{aligned}\tag{56}$$

$$\tilde{\lambda}_F=\lambda_F\left(1-\frac{\mu}{2}\right)+\lambda_B 4\nu L_B^2\left(1-\frac{\mu}{2}\right),\tag{57}$$

$$\tilde{\lambda}_G=\lambda_G(1-p_{\rm s})+\lambda_B 16\nu L_B^2\left(1+\frac{p_{\rm s}}{\mu}\right)+\lambda_F 4p_{\rm s}\left(1+\frac{p_{\rm s}}{\mu}\right),\tag{58}$$

$$\begin{aligned}\tilde{\lambda}_X=&\lambda_A\frac{2\nu^2\mu^2(\omega_{\rm s}+3)\omega\hat{L}^2}{n}\\&+\lambda_B\left(\frac{2L^2}{\nu}+4\nu\left(\frac{4}{\mu}+5\omega_{\rm s}\right)L_A^2+4\nu\left(\frac{4}{\mu}+\frac{5\omega_{\rm s}}{n}\right)L_B^2\right)\end{aligned}\tag{59}$$

$$+\lambda_C\frac{3\hat{L}^2\mu^2(\omega_{\rm s}+3)}{\nu}+\lambda_D 4\left(\frac{1}{\mu}+\omega_{\rm s}\right)+\lambda_E\omega_{\rm s}+\lambda_F 4\left(\frac{1}{\mu}+\frac{\omega_{\rm s}}{n}\right)+\lambda_G\frac{\omega_{\rm s}}{n},\tag{60}$$

$$\tilde{c}=\lambda_A\frac{\omega\nu^2\sigma^2}{nb}+\lambda_B\frac{\nu^2\sigma^2}{nb}+\lambda_C\frac{\nu^2\sigma^2}{b}.\tag{61}$$

Specifically, the term A^k appears in (44); B^k appears in (46); terms involving C^k are collected from inequalities (44) and (45); F^k appears in (46) and (48); the coefficient of E^k is gathered from inequalities (44), (45), (46), (47) and (49); G^k appears in (46), (48) and (50). Then, we collect the constant terms appearing in (44), (46) and (45). Finally, the coefficients associated with X^k appear in (44), (46), (45), (47), (49), (48), and (50).

Substituting (51) into inequality (43) yields

$$\begin{aligned}&\mathbb{E}_k\left[\Psi^{k+1}\right]+\frac{\gamma}{2}\left\|\nabla f(x^k)\right\|^2\\&\leq\delta^k+\left(\tilde{\lambda}_A+\frac{3\gamma}{2}\right)A^k+\left(\tilde{\lambda}_B+\frac{3\gamma}{2}\right)B^k+\tilde{\lambda}_C C^k+\tilde{\lambda}_D D^k\\&\quad+\tilde{\lambda}_E E^k+\tilde{\lambda}_F F^k+\tilde{\lambda}_G G^k-\left(\frac{1}{2\gamma}-\frac{L}{2}-\tilde{\lambda}_X\right)X^k+\tilde{c}.\end{aligned}\tag{62}$$

Let $\hat{\lambda}_A, \ldots, \hat{\lambda}_G$ denote the coefficients of $A^k, \ldots, G^k$ in inequality (62), respectively. Recall the chosen values for $\lambda_A, \cdots, \lambda_G$:

$$\lambda_A=\frac{3\gamma}{2p}\tag{63}$$

$$\lambda_B = \frac{3\gamma}{2\nu} \tag{64}$$

$$\lambda_C = \frac{3\nu\omega\gamma}{np} \tag{65}$$

$$\lambda_D = \gamma\left(\frac{72\mu L_{\max}^2\omega}{np} + \frac{12L_A^2}{\mu}\right) \tag{66}$$

$$\lambda_E = \gamma\left(\frac{336\omega L_{\max}^2}{np}(\mu + p_{\rm s}) + 96L_A^2\left(\frac{1}{p_{\rm s}} + \frac{p_{\rm s}}{\mu^2}\right)\right) \tag{67}$$

$$\lambda_F = \frac{12\gamma L_B^2}{\mu} \tag{68}$$

$$\lambda_G = 96\gamma L_B^2\left(\frac{p_{\rm s}}{\mu^2} + \frac{1}{p_{\rm s}}\right). \tag{69}$$

We now demonstrate that this choice of $\lambda_A, \ldots, \lambda_G$ ensures $\hat{\lambda}_A \leq \lambda_A, \ldots, \hat{\lambda}_G \leq \lambda_G$. We start with $\hat{\lambda}_A$:

$$\hat{\lambda}_A = \tilde{\lambda}_A + \frac{3\gamma}{2} = (1-p)\lambda_A + \frac{3\gamma}{2}.$$

Observe that selecting λ_A as in (63) ensures $p\lambda_A = \frac{3\gamma}{2}$. Therefore, $\hat{\lambda}_A = \lambda_A$. We apply the same logic to the other coefficients. Then,

$$\hat{\lambda}_B = \tilde{\lambda}_B + \frac{3\gamma}{2} = (1-\nu)\lambda_B + \frac{3\gamma}{2} = (1-\nu)\lambda_B + \nu\lambda_B = \lambda_B.$$

Since $\lambda_C = \frac{6\nu\omega\gamma}{2np}$, we obtain

$$\hat{\lambda}_C = (1-\nu)\lambda_C + \frac{2\nu^2\omega}{n}\lambda_A = (1-\nu)\lambda_C + \frac{2\nu^2\omega}{n}\frac{3\gamma}{2p} = (1-\nu)\lambda_C + \frac{6\nu^2\omega\gamma}{2np} = \lambda_C.$$

Due to our choice of λ_D in (66):

$$\begin{aligned}
\hat{\lambda}_D &= \left(1-\frac{\mu}{2}\right)\lambda_D + \frac{6\nu^2\mu^2\omega L_{\max}^2}{n}\lambda_A + 4\nu L_A^2\left(1-\frac{\mu}{2}\right)\lambda_B + \frac{9L_{\max}^2\mu^2}{\nu}\lambda_C \\
&= \left(1-\frac{\mu}{2}\right)\lambda_D + \frac{6\nu^2\mu^2\omega L_{\max}^2}{n}\frac{3\gamma}{2p} + 4\nu L_A^2\left(1-\frac{\mu}{2}\right)\frac{3\gamma}{2\nu} + \frac{9L_{\max}^2\mu^2}{\nu}\frac{3\nu\omega\gamma}{np} \\
&= \left(1-\frac{\mu}{2}\right)\lambda_D + \gamma\left(\frac{9\nu^2\mu^2\omega L_{\max}^2}{np} + 6L_A^2\left(1-\frac{\mu}{2}\right) + \frac{27\mu^2\omega L_{\max}^2}{np}\right) \\
&\leq \left(1-\frac{\mu}{2}\right)\lambda_D + \gamma\left(\frac{36\mu^2\omega L_{\max}^2}{np} + 6L_A^2\right) \\
&= \left(1-\frac{\mu}{2}\right)\lambda_D + \frac{\mu\gamma}{2}\left(\frac{72\mu\omega L_{\max}^2}{np} + \frac{12L_A^2}{\mu}\right) = \lambda_D.
\end{aligned}$$

The inequality uses $\nu^2 \leq 1$ and $1 - \frac{\mu}{2} \leq 1$. Next,

$$\begin{aligned}
\hat{\lambda}_F &= \left(1-\frac{\mu}{2}\right)\lambda_F + 4\nu L_B^2\left(1-\frac{\mu}{2}\right)\lambda_B \\
&= \left(1-\frac{\mu}{2}\right)\lambda_F + 4\nu L_B^2\left(1-\frac{\mu}{2}\right)\frac{3\gamma}{2\nu} \\
&\leq \left(1-\frac{\mu}{2}\right)\lambda_F + 6\gamma L_B^2 = \left(1-\frac{\mu}{2}\right)\lambda_F + \frac{\mu\gamma}{2}\frac{12L_B^2}{\mu} = \lambda_F.
\end{aligned}$$

Then,

$$\hat{\lambda}_G = (1-p_{\rm s})\lambda_G + \underbrace{16\nu L_B^2\left(1+\frac{p_{\rm s}}{\mu}\right)\lambda_B + 4p_{\rm s}\left(1+\frac{p_{\rm s}}{\mu}\right)\lambda_F}_{I}.$$

We consider I separately:

$$\begin{aligned}
I &= 16\nu L_B^2\left(1+\frac{p_{\rm s}}{\mu}\right)\frac{3\gamma}{2\nu}+4p_{\rm s}\left(1+\frac{p_{\rm s}}{\mu}\right)\frac{12\gamma L_B^2}{\mu}\\
&= p_{\rm s}\gamma L_B^2\left(\frac{24}{p_{\rm s}}+\frac{24}{\mu}+\frac{48}{\mu}\left(1+\frac{p_{\rm s}}{\mu}\right)\right)\\
&= p_{\rm s}\gamma L_B^2\left(\frac{24}{p_{\rm s}}+\frac{24}{\mu}+\frac{48}{\mu}+\frac{48p_{\rm s}}{\mu^2}\right)\\
&= p_{\rm s}\gamma L_B^2\left(\frac{72}{\mu}+\frac{48p_{\rm s}}{\mu^2}+\frac{24}{p_{\rm s}}\right)\\
&\le p_{\rm s}\gamma L_B^2\left(\frac{36p_{\rm s}}{\mu^2}+\frac{36}{p_{\rm s}}+\frac{48p_{\rm s}}{\mu^2}+\frac{24}{p_{\rm s}}\right)\\
&= p_{\rm s}\gamma L_B^2\left(\frac{84p_{\rm s}}{\mu^2}+\frac{60}{p_{\rm s}}\right)\\
&\le 96p_{\rm s}\gamma L_B^2\left(\frac{p_{\rm s}}{\mu^2}+\frac{1}{p_{\rm s}}\right)=p_{\rm s}\lambda_G.
\end{aligned}$$

Here, we apply Young's inequality in the form $\frac{2}{\mu}\le\frac{p_{\rm s}}{\mu^2}+\frac{1}{p_{\rm s}}$ and then our choice of λ_G defined in (69). Therefore, we obtain $\hat{\lambda}_G\le\lambda_G$. Then,

$$\begin{aligned}
\hat{\lambda}_E &= (1-p_{\rm s})\lambda_E\\
&\quad+\underbrace{\frac{8\nu^2\mu^2p_{\rm s}\omega L_{\max}^2}{n}\lambda_A+16\nu L_A^2\left(1+\frac{p_{\rm s}}{\mu}\right)\lambda_B+\frac{12L_{\max}^2\mu^2p_{\rm s}}{\nu}\lambda_C+4p_{\rm s}\left(1+\frac{p_{\rm s}}{\mu}\right)\lambda_D}_{I}.
\end{aligned}$$

As before, we consider I separately:

$$\begin{aligned}
I &= \frac{8\nu^2\mu^2p_{\rm s}\omega L_{\max}^2}{n}\frac{3\gamma}{2p}+16\nu L_A^2\left(1+\frac{p_{\rm s}}{\mu}\right)\frac{3\gamma}{2\nu}\\
&\quad+\frac{12L_{\max}^2\mu^2p_{\rm s}}{\nu}\frac{3\nu\omega\gamma}{np}+4p_{\rm s}\left(1+\frac{p_{\rm s}}{\mu}\right)\gamma\left(\frac{72\mu\omega L_{\max}^2}{np}+\frac{12L_A^2}{\mu}\right)\\
&= p_{\rm s}\gamma\left(\frac{12\nu^2\mu^2\omega L_{\max}^2}{np}+\frac{36\mu^2\omega L_{\max}^2}{np}+L_A^2\left(\frac{24}{p_{\rm s}}+\frac{24}{\mu}\right)\right)\\
&\quad+4p_{\rm s}\gamma\left(1+\frac{p_{\rm s}}{\mu}\right)\left(\frac{72\mu\omega L_{\max}^2}{np}+\frac{12L_A^2}{\mu}\right)\\
&= p_{\rm s}\gamma\frac{\omega L_{\max}^2}{np}\left((12\nu^2+36)\mu^2+288\mu+288p_{\rm s}\right)+p_{\rm s}\gamma L_A^2\left(\frac{72}{\mu}+\frac{48p_{\rm s}}{\mu^2}+\frac{24}{p_{\rm s}}\right)\\
&\le p_{\rm s}\gamma\frac{\omega L_{\max}^2}{np}\left(48\mu+288\mu+288p_{\rm s}\right)+p_{\rm s}\gamma L_A^2\left(\frac{36p_{\rm s}}{\mu^2}+\frac{36}{p_{\rm s}}+\frac{48p_{\rm s}}{\mu^2}+\frac{24}{p_{\rm s}}\right)\\
&\le p_{\rm s}\gamma\left(\frac{336\omega L_{\max}^2}{np}(\mu+p_{\rm s})+L_A^2\left(\frac{84p_{\rm s}}{\mu^2}+\frac{60}{p_{\rm s}}\right)\right)\\
&\le p_{\rm s}\gamma\left(\frac{336\omega L_{\max}^2}{np}(\mu+p_{\rm s})+96L_A^2\left(\frac{1}{p_{\rm s}}+\frac{p_{\rm s}}{\mu^2}\right)\right)=p_{\rm s}\lambda_E.
\end{aligned}$$

Here, the lower-order terms are absorbed using the bounds $\mu\le1$ and $\nu\le1$. To obtain the first inequality, we use the same form of Young's inequality as above. Our choice of λ_E in (67) ensures $\hat{\lambda}_E\le\lambda_E$.

Since we have shown $\hat{\lambda}_A\le\lambda_A,\dots,\hat{\lambda}_G\le\lambda_G$, inequality (62) transforms into

$$\mathbb{E}_k\left[\Psi^{k+1}\right]+\frac{\gamma}{2}\left\|\nabla f(x^k)\right\|^2\le\mathbb{E}_k\left[\Psi^k\right]-\underbrace{\left(\frac{1}{2\gamma}-\frac{L}{2}-\tilde{\lambda}_X\right)}_{\hat{\lambda}_X}X^k+\tilde{c}. \tag{70}$$

Next, we require the noise term to be less than $\frac{\gamma\varepsilon}{4}$:

$$\begin{aligned}\tilde{c} &= \frac{\omega\nu^2\sigma^2}{nb}\lambda_A + \frac{\nu^2\sigma^2}{nb}\lambda_B + \nu^2\frac{\sigma^2}{b}\lambda_C \\ &= \frac{\omega\nu^2\sigma^2}{nb}\frac{3\gamma}{2p} + \frac{\nu^2\sigma^2}{nb}\frac{3\gamma}{2\nu} + \nu^2\frac{\sigma^2}{b}\frac{3\nu\omega\gamma}{np} \\ &= \gamma\left(\frac{3\omega\nu^2\sigma^2}{2npb} + \frac{3\nu\sigma^2}{2nb} + \frac{3\nu^3\omega\sigma^2}{npb}\right) \\ &\leq \frac{\gamma}{2}\left(\frac{9\nu^2\omega\sigma^2}{npb} + \frac{3\nu\sigma^2}{nb}\right) \leq \frac{\gamma\varepsilon}{4}.\end{aligned} \tag{71}$$

The last inequality holds due to our choice of the momentum parameter ν. Finally, to establish convergence, we choose γ such that $\hat{\lambda}_X \geq 0$.

$$\begin{aligned}\hat{\lambda}_X = \frac{1}{2\gamma} - \frac{L}{2} - \frac{2\nu^2\mu^2(\omega_{\mathrm{s}}+3)\omega\hat{L}^2}{n}\lambda_A - \left(\frac{2L^2}{\nu} + 4\nu\left(\frac{4}{\mu}+5\omega_{\mathrm{s}}\right)L_A^2 + 4\nu\left(\frac{4}{\mu}+\frac{5\omega_{\mathrm{s}}}{n}\right)L_B^2\right)\lambda_B \\ - \frac{3\hat{L}^2\mu^2(\omega_{\mathrm{s}}+3)}{\nu}\lambda_C - 4\left(\frac{1}{\mu}+\omega_{\mathrm{s}}\right)\lambda_D - 4\left(\frac{1}{\mu}+\frac{\omega_{\mathrm{s}}}{n}\right)\lambda_F \\ - \omega_{\mathrm{s}}\lambda_E - \frac{\omega_{\mathrm{s}}}{n}\lambda_G \geq 0.\end{aligned} \tag{72}$$

Substituting the values of $\lambda_A, \cdots, \lambda_G$ from (63)–(69) into (72), we obtain $\hat{\lambda}_X = \frac{1}{2\gamma} - \frac{L}{2} - \tilde{A}\gamma$, where $\tilde{A}$ is the aggregate coefficient. Using $L \leq L_{\max}$, we bound it as:

$$\tilde{A} \leq C_{\hat{L}}\hat{L}^2 + C_{\max}L_{\max}^2 + C_A L_A^2 + C_B L_B^2.$$

By grouping terms, we derive each coefficient. We start with $C_{\hat{L}}$:

$$\begin{aligned}C_{\hat{L}} &= \frac{1}{\gamma}\left(\frac{2\nu^2\mu^2(\omega_{\mathrm{s}}+3)\omega}{n}\lambda_A + \frac{3\mu^2(\omega_{\mathrm{s}}+3)}{\nu}\lambda_C\right) \\ &= \frac{2\nu^2\mu^2(\omega_{\mathrm{s}}+3)\omega}{n}\frac{3}{2p} + \frac{3\mu^2(\omega_{\mathrm{s}}+3)}{\nu}\frac{3\nu\omega}{np} \\ &= \frac{3\mu^2\omega(\omega_{\mathrm{s}}+3)}{np}(\nu^2+3) \leq \frac{36\mu^2\omega(\omega_{\mathrm{s}}+1)}{np}.\end{aligned} \tag{73}$$

Next, we derive the coefficient for L_A, which appears not only in (72) but also in the expressions for λ_D (66) and λ_E (67).

$$\begin{aligned}C_A &= 6\left(\frac{4}{\mu}+5\omega_{\mathrm{s}}\right) + 4\left(\frac{1}{\mu}+\omega_{\mathrm{s}}\right)\frac{12}{\mu} + 96\omega_{\mathrm{s}}\left(\frac{p_{\mathrm{s}}}{\mu^2}+\frac{1}{p_{\mathrm{s}}}\right) \\ &= \frac{24}{\mu} + 30\omega_{\mathrm{s}} + \frac{48}{\mu^2} + \frac{48\omega_{\mathrm{s}}}{\mu} + \frac{96\omega_{\mathrm{s}}p_{\mathrm{s}}}{\mu^2} + \frac{96\omega_{\mathrm{s}}}{p_{\mathrm{s}}} \\ &= \omega_{\mathrm{s}}\left(\frac{48}{\mu} + 30 + \frac{96p_{\mathrm{s}}}{\mu^2} + \frac{96}{p_{\mathrm{s}}}\right) + \frac{48}{\mu^2} + \frac{24}{\mu} \\ &\leq \omega_{\mathrm{s}}\left(\frac{48}{\mu} + \frac{30}{p_{\mathrm{s}}} + \frac{96p_{\mathrm{s}}}{\mu^2} + \frac{96}{p_{\mathrm{s}}}\right) + \frac{72}{\mu^2} \\ &\leq \omega_{\mathrm{s}}\left(\frac{24p_{\mathrm{s}}}{\mu^2} + \frac{24}{p_{\mathrm{s}}} + \frac{30}{p_{\mathrm{s}}} + \frac{96p_{\mathrm{s}}}{\mu^2} + \frac{96}{p_{\mathrm{s}}}\right) + \frac{72}{\mu^2} \\ &= \omega_{\mathrm{s}}\left(\frac{120p_{\mathrm{s}}}{\mu^2} + \frac{150}{p_{\mathrm{s}}}\right) + \frac{72}{\mu^2} \\ &\leq 150\omega_{\mathrm{s}}\left(\frac{p_{\mathrm{s}}}{\mu^2} + \frac{1}{p_{\mathrm{s}}}\right) + \frac{72}{\mu^2}.\end{aligned} \tag{74}$$

Next, we derive the coefficient for L_B, which appears in (72) and in the expressions for λ_F (68) and λ_G (69):

$$\begin{aligned} C_B &= 6\left(\frac{4}{\mu}+\frac{5\omega_{\mathrm{s}}}{n}\right)+4\left(\frac{1}{\mu}+\frac{\omega_{\mathrm{s}}}{n}\right)\frac{12}{\mu}+96\frac{\omega_{\mathrm{s}}}{n}\left(\frac{p_{\mathrm{s}}}{\mu^2}+\frac{1}{p_{\mathrm{s}}}\right) \\ &\leq 150\frac{\omega_{\mathrm{s}}}{n}\left(\frac{p_{\mathrm{s}}}{\mu^2}+\frac{1}{p_{\mathrm{s}}}\right)+\frac{72}{\mu^2}. \end{aligned} \tag{75}$$

We obtain the last inequality similarly to C_A. It remains to determine $C_{\max}$. Since $L_{\max}^2$ appears in λ_D (66) and λ_E (67), and the global-gradient term contributes $3L^2/\nu^2 \leq 3L_{\max}^2/\nu^2$, substituting them into (72) yields

$$\begin{aligned} C_{\max} &= \frac{3}{\nu^2}+4\left(\frac{1}{\mu}+\omega_{\mathrm{s}}\right)\frac{72\mu\omega}{np}+\omega_{\mathrm{s}}\frac{336\omega}{np}(\mu+p_{\mathrm{s}}) \\ &= \frac{3}{\nu^2}+\frac{\omega}{np}\left(288(1+\mu\omega_{\mathrm{s}})+336(\mu+p_{\mathrm{s}})\omega_{\mathrm{s}}\right) \\ &\leq \frac{3}{\nu^2}+\frac{336\omega}{np}\left(1+\mu\omega_{\mathrm{s}}+\mu\omega_{\mathrm{s}}+p_{\mathrm{s}}\omega_{\mathrm{s}}\right) \\ &\leq \frac{3}{\nu^2}+\frac{672\omega}{np}\left(1+(\mu+p_{\mathrm{s}})\omega_{\mathrm{s}}\right). \end{aligned} \tag{76}$$

Lemma D.5 implies that the condition $\hat{\lambda}_X \geq 0$ holds for any γ less than

$$\tilde{\gamma} := \left(L+\sqrt{2\left(C_{\hat{L}}\hat{L}^2+C_AL_A^2+C_BL_B^2+C_{\max}L_{\max}^2\right)}\right)^{-1}. \tag{77}$$

Combining (73), (74), (75) and (76), we obtain a lower bound on $\tilde{\gamma}$:

$$\begin{aligned} &C_AL_A^2+C_BL_B^2+C_{\max}L_{\max}^2+C_{\hat{L}}\hat{L}^2 \\ &\leq 150\omega_{\mathrm{s}}\left(\frac{p_{\mathrm{s}}}{\mu^2}+\frac{1}{p_{\mathrm{s}}}\right)L_A^2+150\frac{\omega_{\mathrm{s}}}{n}\left(\frac{p_{\mathrm{s}}}{\mu^2}+\frac{1}{p_{\mathrm{s}}}\right)L_B^2 \\ &\quad+\frac{72}{\mu^2}\left(L_B^2+L_A^2\right)+\frac{3L_{\max}^2}{\nu^2} \\ &\quad+672\frac{\omega}{np}\left(1+(\mu+p_{\mathrm{s}})\omega_{\mathrm{s}}\right)L_{\max}^2+36\frac{\mu^2\omega(\omega_{\mathrm{s}}+1)}{np}\hat{L}^2. \end{aligned} \tag{78}$$

Consider the terms involving $L_{\max}$ and $\hat{L}$ separately. Using $\mu \leq 1$ and $\hat{L}^2 \leq L_{\max}^2$, we get

$$C_{\max}L_{\max}^2+C_{\hat{L}}\hat{L}^2 \leq \left(\frac{3}{\nu^2}+672\frac{\omega}{np}\left(1+(\mu+p_{\mathrm{s}})\omega_{\mathrm{s}}\right)+36\frac{\mu^2\omega(\omega_{\mathrm{s}}+1)}{np}\right)L_{\max}^2 \tag{79}$$

$$\leq \left(\frac{3}{\nu^2}+672\frac{\omega(1+p_{\mathrm{s}}\omega_{\mathrm{s}})}{np}+672\frac{\mu\omega\omega_{\mathrm{s}}}{np}+36\frac{\mu\omega\omega_{\mathrm{s}}}{np}+36\frac{\omega}{np}\right)L_{\max}^2 \tag{80}$$

$$\leq 708\left(\frac{\mu\omega\omega_{\mathrm{s}}}{np}+\frac{\omega(1+\omega_sp_{\mathrm{s}})}{np}+\frac{1}{\nu^2}\right)L_{\max}^2. \tag{81}$$

Thus, substituting (79) into (78) yields

$$\begin{aligned} &C_{\hat{L}}\hat{L}^2+C_AL_A^2+C_BL_B^2+C_{\max}L_{\max}^2 \\ &\leq 708\left(\tilde{C}_AL_A^2+\tilde{C}_BL_B^2+\tilde{C}_{\max}L_{\max}^2\right) \\ &= 708\underbrace{\left(\frac{\omega_{\mathrm{s}}p_{\mathrm{s}}+1}{\mu^2}+\frac{\omega_{\mathrm{s}}}{p_{\mathrm{s}}}\right)}_{\tilde{C}_A}L_A^2+708\underbrace{\left(\frac{\frac{\omega_{\mathrm{s}}}{n}p_{\mathrm{s}}+1}{\mu^2}+\frac{\omega_{\mathrm{s}}}{np_{\mathrm{s}}}\right)}_{\tilde{C}_B}L_B^2 \end{aligned} \tag{82}$$

$$+ 708 \underbrace{\left(\frac{\mu\omega\omega_{\mathrm{s}}}{np} + \frac{\omega(1+\omega_s p_{\mathrm{s}})}{np} + \frac{1}{\nu^2} \right)}_{\tilde{C}_{\max}} L_{\max}^2.$$

Thus, combining (77) and (82), we can take γ such that

$$\gamma = \left(L + \sqrt{1416 \left(\tilde{C}_A L_A^2 + \tilde{C}_B L_B^2 + \tilde{C}_{\max} L_{\max}^2 \right)} \right)^{-1}.$$

Note that $\gamma \leq \tilde{\gamma}$, implying $\hat{\lambda}_X \geq 0$ holds. Recall that our choice of ν ensures $\tilde{c} \leq \frac{\gamma\varepsilon}{4}$ in (71). Combining these results, inequality (70) transforms into

$$\mathbb{E}_k \left[\Psi^{k+1} \right] + \frac{\gamma}{2} \left\| \nabla f(x^k) \right\|^2 \leq \mathbb{E}_k \left[\Psi^k \right] + \frac{\gamma\varepsilon}{4}.$$

Averaging over $k = 0, \ldots, K-1$ and taking the full expectation, we obtain

$$\frac{\mathbb{E}\left[\Psi^K\right]}{K} + \frac{\gamma}{2} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\left[\left\| \nabla f(x^k) \right\|^2 \right] \leq \frac{\mathbb{E}\left[\Psi^0\right]}{K} + \frac{\gamma\varepsilon}{4}.$$

Dropping the non-negative term $\mathbb{E}\left[\Psi^K\right]$, dividing both sides of the inequality by $\frac{\gamma}{2}$, we derive

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\left[\left\| \nabla f(x^k) \right\|^2 \right] \leq \frac{2\mathbb{E}\left[\Psi^0\right]}{\gamma K} + \frac{\varepsilon}{2}.$$

□

Theorem F.10 (Equal momentum coefficients)**.** *Let the assumptions of Theorem F.9 be satisfied. Assume that* $\nu = \mu = \eta := \min\left\{ \frac{1}{6}\sqrt{\frac{bn\varepsilon}{\omega(\omega+1)\sigma^2}}, \frac{bn\varepsilon}{12\sigma^2}, \left(\frac{n}{\omega(\omega+1)\omega_{\mathrm{s}}}\right)^{\frac{1}{3}}, 1 \right\}$, $p_{\mathrm{s}} = \frac{1}{\omega_{\mathrm{s}}+1}, p = \frac{1}{\omega+1}$. *Then, for any*

$$\gamma \leq \left(4\sqrt{c \left(\omega_{\mathrm{s}}(\omega_{\mathrm{s}}+1) L_A^2 + \frac{\omega_{\mathrm{s}}}{n}(\omega_{\mathrm{s}}+1) L_B^2 + \left(\frac{\omega(\omega+1)}{n} + \frac{1}{\eta^2} \right) L_{\max}^2 \right)} \right)^{-1}, \tag{83}$$

where c *is the multiplicative factor defined in Theorem F.9, the following holds:*

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\left[\left\| \nabla f(x^k) \right\|^2 \right] \leq \frac{2\mathbb{E}\left[\Psi^0\right]}{\gamma K} + \frac{\varepsilon}{2}.$$

Proof. For our choice of ν and $p = \frac{1}{\omega+1}$, we have

$$\nu^2 \omega(\omega+1)\sigma^2 \leq \frac{bn\varepsilon}{36}, \qquad \nu\sigma^2 \leq \frac{bn\varepsilon}{12}.$$

Consequently,

$$\begin{aligned} \frac{9\nu^2\omega\sigma^2}{npb} + \frac{3\nu\sigma^2}{nb} &= \frac{9}{nb}\nu^2\omega(\omega+1)\sigma^2 + \frac{3}{nb}\nu\sigma^2 \\ &\leq \frac{9}{nb}\frac{bn\varepsilon}{36} + \frac{3}{nb}\frac{bn\varepsilon}{12} = \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2}. \end{aligned}$$

Therefore, the condition of Theorem F.9 is satisfied.

Substituting $\mu = \nu = \eta$ into the expression for the step size (40) from Theorem F.9 yields

$$\begin{aligned} &\tilde{C}_A L_A^2 + \tilde{C}_B L_B^2 + \tilde{C}_{\max} L_{\max}^2 \\ &= \left(\frac{\omega_{\mathrm{s}} p_{\mathrm{s}} + 1}{\eta^2} + \frac{\omega_{\mathrm{s}}}{p_{\mathrm{s}}} \right) L_A^2 + \left(\frac{\frac{\omega_{\mathrm{s}}}{n} p_{\mathrm{s}} + 1}{\eta^2} + \frac{\omega_{\mathrm{s}}}{np_{\mathrm{s}}} \right) L_B^2 \end{aligned} \tag{84}$$

$$
\begin{aligned}
&+\left(\frac{\eta\omega\omega_{\rm s}}{np}+\frac{\omega(1+\omega_{\rm s}p_{\rm s})}{np}+\frac{1}{\eta^2}\right)L_{\max}^2\\
&=\left(\omega_{\rm s}\left(\frac{p_{\rm s}}{\eta^2}+\frac{1}{p_{\rm s}}\right)+\frac{1}{\eta^2}\right)L_A^2+\left(\frac{\omega_{\rm s}}{n}\left(\frac{p_{\rm s}}{\eta^2}+\frac{1}{p_{\rm s}}\right)+\frac{1}{\eta^2}\right)L_B^2\\
&+\left(\frac{\eta\omega\omega_{\rm s}}{np}+\frac{\omega(1+\omega_{\rm s}p_{\rm s})}{np}+\frac{1}{\eta^2}\right)L_{\max}^2.
\end{aligned}
$$

Substituting our choice of probabilities into (84), we obtain

$$
\begin{aligned}
&\tilde{C}_AL_A^2+\tilde{C}_BL_B^2+\tilde{C}_{\max}L_{\max}^2\\
&=\left(\omega_{\rm s}(\omega_{\rm s}+1)+\frac{1+\frac{\omega_{\rm s}}{\omega_{\rm s}+1}}{\eta^2}\right)L_A^2+\left(\frac{\omega_{\rm s}(\omega_{\rm s}+1)}{n}+\frac{1+\frac{\omega_{\rm s}}{n(\omega_{\rm s}+1)}}{\eta^2}\right)L_B^2\\
&\quad+\left(\frac{\eta\omega(\omega+1)\omega_{\rm s}}{n}+\frac{\omega(\omega+1)}{n}\left(1+\frac{\omega_{\rm s}}{\omega_{\rm s}+1}\right)+\frac{1}{\eta^2}\right)L_{\max}^2\\
&\le\left(\omega_{\rm s}(\omega_{\rm s}+1)+\frac{2}{\eta^2}\right)L_A^2+\left(\frac{\omega_{\rm s}(\omega_{\rm s}+1)}{n}+\frac{1+\frac{1}{n}}{\eta^2}\right)L_B^2\\
&\quad+\left(\frac{\eta\omega(\omega+1)\omega_{\rm s}}{n}+\frac{2\omega(\omega+1)}{n}+\frac{1}{\eta^2}\right)L_{\max}^2.
\end{aligned}
$$

Since $\frac{2}{\eta^2}L_A^2+\frac{1+\frac{1}{n}}{\eta^2}L_B^2\le\frac{2}{\eta^2}\left(L_A^2+L_B^2\right)\le\frac{4}{\eta^2}L_{\max}^2$, we get

$$
\begin{aligned}
&\tilde{C}_AL_A^2+\tilde{C}_BL_B^2+\tilde{C}_{\max}L_{\max}^2\\
&\le\omega_{\rm s}(\omega_{\rm s}+1)L_A^2+\frac{\omega_{\rm s}(\omega_{\rm s}+1)}{n}L_B^2+\left(\frac{\eta\omega(\omega+1)\omega_{\rm s}}{n}+\frac{2\omega(\omega+1)}{n}+\frac{5}{\eta^2}\right)L_{\max}^2.
\end{aligned}
$$

Using our choice of η, we ensure

$$
\eta\le\left(\frac{n}{\omega(\omega+1)\omega_{\rm s}}\right)^{\frac{1}{3}}\quad\Longrightarrow\quad\frac{\eta\omega(\omega+1)\omega_{\rm s}}{n}\le\frac{1}{\eta^2}.
$$

Thus,

$$
\begin{aligned}
&\tilde{C}_AL_A^2+\tilde{C}_BL_B^2+\tilde{C}_{\max}L_{\max}^2\\
&\le\omega_{\rm s}(\omega_{\rm s}+1)L_A^2+\frac{\omega_{\rm s}(\omega_{\rm s}+1)}{n}L_B^2+\left(\frac{2\omega(\omega+1)}{n}+\frac{6}{\eta^2}\right)L_{\max}^2.
\end{aligned}
$$

Since $L\le\frac{L_{\max}}{\eta}$, applying the inequality $\sqrt{a}+\sqrt{b}\le\sqrt{2(a+b)}$, we obtain

$$
\begin{aligned}
&L+\sqrt{c\left(\tilde{C}_AL_A^2+\tilde{C}_BL_B^2+\tilde{C}_{\max}L_{\max}^2\right)}\\
&\le\sqrt{2L^2+2c\left(\tilde{C}_AL_A^2+\tilde{C}_BL_B^2+\tilde{C}_{\max}L_{\max}^2\right)}\\
&\le\left(\frac{2L_{\max}^2}{\eta^2}+2c\left(\omega_{\rm s}(\omega_{\rm s}+1)L_A^2+\frac{\omega_{\rm s}(\omega_{\rm s}+1)}{n}L_B^2+\left(\frac{2\omega(\omega+1)}{n}+\frac{6}{\eta^2}\right)L_{\max}^2\right)\right)^{\frac{1}{2}}\\
&\le\left(2c\left(\omega_{\rm s}(\omega_{\rm s}+1)L_A^2+\frac{\omega_{\rm s}(\omega_{\rm s}+1)}{n}L_B^2+\left(\frac{2\omega(\omega+1)}{n}+\frac{7}{\eta^2}\right)L_{\max}^2\right)\right)^{\frac{1}{2}}\\
&\le 4\sqrt{c\left(\omega_{\rm s}(\omega_{\rm s}+1)L_A^2+\frac{\omega_{\rm s}(\omega_{\rm s}+1)}{n}L_B^2+\left(\frac{\omega(\omega+1)}{n}+\frac{1}{\eta^2}\right)L_{\max}^2\right)}.
\end{aligned}
$$

Therefore, the step size in (83) satisfies (40). □

Theorem 3.4 (Iteration complexity). *Let Assumptions 3.1, 1.2, 3.2, and 3.3 be satisfied. Then, M4 with momentum parameters* $\nu=\mu=\eta:=\min\left\{\frac{1}{6}\sqrt{\frac{bn\varepsilon}{\omega(\omega+1)\sigma^2}},\frac{bn\varepsilon}{12\sigma^2},\left(\frac{n}{\omega(\omega+1)\omega_{\rm s}}\right)^{\frac{1}{3}},1\right\}$,

$p_{\rm s} = \frac{1}{\omega_{\rm s}+1}, p = \frac{1}{\omega+1}$, $v_i^0 = \frac{1}{b_{\rm init}}\sum_{b=1}^{b_{\rm init}} \nabla f_i(x_i^0;\xi_{i,b}^0)$ *where* $b_{\rm init} = \left\lceil \frac{b}{\nu} \right\rceil$, *and step size* $\gamma = \Theta\left(\sqrt{\left(\omega_{\rm s}(\omega_{\rm s}+1)L_A^2 + \frac{\omega_{\rm s}}{n}(\omega_{\rm s}+1)L_B^2 + \left(\frac{\omega(\omega+1)}{n} + \frac{1}{\eta^2}\right)L_{\max}^2\right)}\right)^{-1}$ *finds an* ε*–stationary point of* (12) *after at most*

$$K = \mathcal{O}\left(\frac{\Delta}{\gamma\varepsilon} + \max\left\{\sqrt{\frac{\omega(\omega+1)\sigma^2}{bn\varepsilon}}, \frac{\sigma^2}{bn\varepsilon}, \left(\frac{\omega(\omega+1)\omega_{\rm s}}{n}\right)^{\frac{1}{3}}, 1\right\}\right)$$

iterations.

Proof. Theorem F.10 yields that $K = \mathcal{O}\left(\frac{\mathbb{E}\left[\Psi^0\right]}{\gamma\varepsilon}\right) = \mathcal{O}\left(\frac{\Delta}{\gamma\varepsilon} + \frac{\mathbb{E}\left[\Psi^0-\Delta\right]}{\gamma\varepsilon}\right)$. Consider the second term separately. Since $g^0 = v^0$ and $x^0 = w_i^0 = x_i^0$, we get

$$\begin{aligned}
\mathbb{E}\left[\Psi^0 - \Delta\right] &= \delta^0 - \Delta + \lambda_A\mathbb{E}\left[\left\|g^0 - v^0\right\|^2\right] \\
&\quad + \lambda_B\mathbb{E}\left[\left\|v^0 - \nabla f(x^0)\right\|^2\right] + \lambda_C\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|v_i^0 - \nabla f_i(x_i^0)\right\|^2\right] \\
&\quad + \lambda_D\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|w_i^0 - x_i^0\right\|^2\right] + \lambda_E\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|w_i^0 - x^0\right\|^2\right] \\
&\quad + \lambda_F\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n(w_i^0 - x_i^0)\right\|^2\right] + \lambda_G\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n w_i^0 - x^0\right\|^2\right] \\
&\leq \lambda_B\mathbb{E}\left[\left\|v^0 - \nabla f(x^0)\right\|^2\right] + \lambda_C\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|v_i^0 - \nabla f_i(x_i^0)\right\|^2\right].
\end{aligned}$$

Here we drop the term $\delta^0 - \Delta$, as it is non-positive under Assumption 1.2. Recall our choice of λ_B in (64) and λ_C in (65).

$$\begin{aligned}
\frac{\mathbb{E}\left[\Psi^0 - \Delta\right]}{\gamma\varepsilon} &\leq \frac{1}{\gamma\varepsilon}\left(\frac{3\gamma}{2\nu}\mathbb{E}\left[\left\|v^0 - \nabla f(x^0)\right\|^2\right] + \frac{3\nu\omega\gamma}{np}\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|v_i^0 - \nabla f_i(x_i^0)\right\|^2\right]\right) \\
&= \frac{3}{2\nu\varepsilon}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\frac{1}{b_{\rm init}}\sum_{r=1}^{b_{\rm init}}\nabla f_i(x_i^0;\xi_{i,r}^0) - \frac{1}{n}\sum_{i=1}^n\nabla f_i(x_i^0)\right\|^2\right] \\
&\quad + \frac{3\nu\omega}{np\varepsilon}\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\frac{1}{b_{\rm init}}\sum_{r=1}^{b_{\rm init}}\nabla f_i(x_i^0;\xi_{i,r}^0) - \nabla f_i(x_i^0)\right\|^2\right] \\
&= \frac{3}{2\nu\varepsilon}\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}\left[\left\|\frac{1}{b_{\rm init}}\sum_{r=1}^{b_{\rm init}}\left(\nabla f_i(x_i^0;\xi_{i,r}^0) - \nabla f_i(x_i^0)\right)\right\|^2\right] \\
&\quad + \frac{3\nu\omega}{np\varepsilon}\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\frac{1}{b_{\rm init}}\sum_{r=1}^{b_{\rm init}}\left(\nabla f_i(x_i^0;\xi_{i,r}^0) - \nabla f_i(x_i^0)\right)\right\|^2\right] \\
&= \frac{3}{2\nu\varepsilon n^2 b_{\rm init}^2}\sum_{i=1}^n\sum_{r=1}^{b_{\rm init}}\mathbb{E}\left[\left\|\nabla f_i(x_i^0;\xi_{i,r}^0) - \nabla f_i(x_i^0)\right\|^2\right] \\
&\quad + \frac{3\nu\omega}{n^2 p\varepsilon b_{\rm init}^2}\sum_{i=1}^n\sum_{r=1}^{b_{\rm init}}\mathbb{E}\left[\left\|\nabla f_i(x_i^0;\xi_{i,r}^0) - \nabla f_i(x_i^0)\right\|^2\right].
\end{aligned}$$

Here we use $x_i^0 = x^0$, unbiasedness, and independence of the initial samples across workers and minibatch elements. Using Assumption 3.2, we get

$$\frac{\mathbb{E}\left[\Psi^0 - \Delta\right]}{\gamma\varepsilon} \leq \frac{3\sigma^2}{2\nu n b_{\rm init}\varepsilon} + \frac{3\nu\omega\sigma^2}{npb_{\rm init}\varepsilon} = \frac{1}{\nu b_{\rm init}}\left(\frac{3\nu^2\omega\sigma^2}{np\varepsilon} + \frac{3\sigma^2}{2n\varepsilon}\right). \tag{85}$$

Our choice of ν implies $\frac{\nu^2\omega\sigma^2}{np\varepsilon} \leq \frac{b}{36}$ and $\frac{\nu\sigma^2}{n\varepsilon} \leq \frac{b}{12}$. Substituting these bounds into (85) yields

$$\begin{aligned}\frac{\mathbb{E}\left[\Psi^0 - \Delta\right]}{\gamma\varepsilon} &\leq \frac{1}{\nu b_{\text{init}}}\left(\frac{3b}{36} + \frac{3b}{24\nu}\right) \\ &= \frac{b}{12\nu b_{\text{init}}} + \frac{b}{8\nu^2 b_{\text{init}}} \leq \frac{b}{12\nu^2 b_{\text{init}}} + \frac{b}{8\nu^2 b_{\text{init}}} = \frac{5b}{24\nu^2 b_{\text{init}}}.\end{aligned} \tag{86}$$

The last inequality uses $\nu \leq 1$. Plugging in our choice of b_{init}, we obtain

$$\frac{\mathbb{E}\left[\Psi^0 - \Delta\right]}{\gamma\varepsilon} \leq \frac{5b}{24\nu^2\left\lceil\frac{b}{\nu}\right\rceil} \leq \frac{5b}{24\nu^2\frac{b}{\nu}} = \frac{5}{24\nu}.$$

Therefore, using our choice of $\nu = \eta$,

$$K = \mathcal{O}\left(\frac{\Delta}{\gamma\varepsilon} + \frac{1}{\nu}\right) = \mathcal{O}\left(\frac{\Delta}{\gamma\varepsilon} + \max\left\{\sqrt{\frac{\omega(\omega+1)\sigma^2}{bn\varepsilon}}, \frac{\sigma^2}{bn\varepsilon}, \left(\frac{\omega(\omega+1)\omega_{\text{s}}}{n}\right)^{\frac{1}{3}}, 1\right\}\right).$$

□

Remark F.11 (Intuition behind b_{init}). We select b_{init} such that the additional iterations incurred by small initial batch size do not outweigh the computational cost of using a large b_{init}. This requirement can be formalized as

$$\underbrace{\frac{\mathbb{E}\left[\Psi^0 - \Delta\right]}{\gamma\varepsilon}}_{\text{extra iterations}} \times bh = \mathcal{O}\left(b_{\text{init}}h\right) \iff \frac{\mathbb{E}\left[\Psi^0 - \Delta\right]}{\gamma\varepsilon} \times b = \mathcal{O}\left(b_{\text{init}}\right).$$

From (86), we have $\frac{\mathbb{E}\left[\Psi^0-\Delta\right]}{\gamma\varepsilon} = \mathcal{O}\left(\frac{b}{\nu^2 b_{\text{init}}}\right)$, hence it is sufficient to require $b_{\text{init}} \geq \frac{b^2}{\nu^2 b_{\text{init}}}$. Consequently,

$$b_{\text{init}}^2 \geq \frac{b^2}{\nu^2} \iff b_{\text{init}} \geq \frac{b}{\nu}. \tag{87}$$

Thus, this choice of b_{init} ensures $\frac{\mathbb{E}\left[\Psi^0-\Delta\right]}{\gamma\varepsilon} \times b = \mathcal{O}\left(b_{\text{init}}\right)$. For $b_{\text{init}} = \left\lceil\frac{b}{\nu}\right\rceil$, we also have

$$b_{\text{init}}h \leq \left(\frac{b}{\nu} + 1\right)h \leq \frac{2bh}{\nu},$$

since $b \geq 1$ and $\nu \leq 1$.

F.4 Time complexity

Theorem 3.5 (Time Complexity)**.** *Consider the assumptions and result of Theorem 3.4. Additionally, assume that Assumption 1.4 is satisfied and the workers and server use RandK compressor with* $K = \min\{d, \left\lfloor\frac{t}{\tau}\right\rfloor\}$ *and* $K = \min\{d, \left\lfloor\frac{t}{\kappa}\right\rfloor\}$*, respectively, and use batch size* $b = \left\lfloor\frac{t}{h}\right\rfloor$*, where* $t = \max\left\{h, \tau, \kappa, \sqrt[3]{d^2\tau^2 h\sigma^2/n\varepsilon}\right\}$*. Then, the expected time for M4 to find an* ε*–stationary point of* (12) *is at most*

$$\mathcal{O}\left(\max\left\{h, \tau, \kappa, \frac{d\kappa}{\sqrt{n}}, \frac{d\tau}{\sqrt{n}}, \left(\frac{d^3\tau^2\kappa}{n}\right)^{\frac{1}{3}}, \left(\frac{d^2\tau^2 h\sigma^2}{n\varepsilon}\right)^{\frac{1}{3}}, \frac{\sigma^2 h}{n\varepsilon}\right\}\frac{L_{\max}\Delta}{\varepsilon} + \frac{d\kappa L_A\Delta}{\varepsilon} + d\tau\right) \tag{14}$$

seconds, including the initial computation of minibatches of size b_{init} *and the initial transmission of the vectors* v_i^0 *to the server.*

Proof. Note that the time of each iteration is almost the same as in Section E.1 and consists of sending 2ℓ coordinates on average by the server, sending $2m$ coordinates on average by the workers

and computing b gradients. To ensure $v^0 = g^0$, the workers send the non-compressed vectors v_i^0 to the server before the first iteration, which takes at most $d\tau$ seconds. Consequently,

$$\mathbb{E}\left[T_{\text{time}}\right] \leq \sum_{k=0}^{K-1} \mathbb{E}\left[t_{\text{iter}}^k\right] + b_{\text{init}}h + d\tau \leq \sum_{k=0}^{K-1} 5t + b_{\text{init}}h + d\tau = 5t \times K + b_{\text{init}}h + d\tau.$$

Since at least one gradient is computed and compressed by each worker, and the server sends a new point compressed by at least one compressor, $t \geq \max\{h, \tau, \kappa\}$.

Theorem 3.4 yields

$$K = \mathcal{O}\left(\frac{\Delta}{\gamma\varepsilon} + \max\left\{\sqrt{\frac{\omega(\omega+1)\sigma^2}{bn\varepsilon}}, \frac{\sigma^2}{bn\varepsilon}, \left(\frac{\omega(\omega+1)\omega_{\text{s}}}{n}\right)^{\frac{1}{3}}, 1\right\}\right).$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[T_{\text{time}}\right] &= \mathcal{O}\left(\frac{t\Delta}{\gamma\varepsilon} + t \times \max\left\{\sqrt{\frac{\omega(\omega+1)\sigma^2}{bn\varepsilon}}, \frac{\sigma^2}{bn\varepsilon}, \left(\frac{\omega(\omega+1)\omega_{\text{s}}}{n}\right)^{\frac{1}{3}}, 1\right\} + b_{\text{init}}h + d\tau\right)\\ &= \mathcal{O}\left(\frac{t\Delta}{\gamma\varepsilon} + \frac{t}{\nu} + b_{\text{init}}h + d\tau\right).\end{aligned}$$

Since $b = \left\lfloor \frac{t}{h} \right\rfloor \geq \frac{t}{2h}$ and $t \leq 2bh$, our choice of b_{init} yields $\frac{t}{\nu} \leq \frac{2bh}{\nu} \leq 2\left\lceil \frac{b}{\nu} \right\rceil h = 2b_{\text{init}}h$. Thus, $\mathbb{E}\left[T_{\text{time}}\right] = \mathcal{O}\left(\frac{t\Delta}{\gamma\varepsilon} + b_{\text{init}}h + d\tau\right)$. Consider $b_{\text{init}}h$ separately. Since $\nu = \eta$, $b \geq 1$, and $\eta \leq 1$, we get

$$b_{\text{init}}h = h\left\lceil \frac{b}{\eta} \right\rceil \leq h\left(\frac{b}{\eta} + 1\right) \leq \frac{2bh}{\eta} \leq \frac{2t}{\eta}.$$

Consequently, $\mathbb{E}\left[T_{\text{time}}\right] = \mathcal{O}\left(\frac{t\Delta}{\gamma\varepsilon} + \frac{t}{\eta} + d\tau\right)$. Substituting γ defined in (83) into the time complexity, we obtain

$$\begin{aligned}\mathbb{E}\left[T_{\text{time}}\right] &= \mathcal{O}\left(\frac{t\Delta}{\varepsilon}\left(\sqrt{\omega_{\text{s}}(\omega_{\text{s}}+1)L_A^2 + \frac{\omega_{\text{s}}}{n}(\omega_{\text{s}}+1)L_B^2 + \left(\frac{\omega(\omega+1)}{n} + \frac{1}{\eta^2}\right)L_{\max}^2}\right) + \frac{t}{\eta} + d\tau\right)\\ &= \mathcal{O}\left(\frac{t\Delta}{\varepsilon}\left(\sqrt{\omega_{\text{s}}(\omega_{\text{s}}+1)}L_A + \sqrt{\frac{\omega_{\text{s}}}{n}(\omega_{\text{s}}+1)}L_B + \left(\sqrt{\frac{\omega(\omega+1)}{n}} + \frac{1}{\eta}\right)L_{\max}\right) + \frac{t}{\eta} + d\tau\right).\end{aligned}$$

Recall that the compressors' parameters are $\omega = \frac{d}{m} - 1 \leq \frac{d}{m}$ and $\omega_{\text{s}} = \frac{d}{\ell} - 1 \leq \frac{d}{\ell}$, where $m = \min\left\{d, \left\lfloor \frac{t}{\tau} \right\rfloor\right\}$ and $\ell = \min\left\{d, \left\lfloor \frac{t}{\kappa} \right\rfloor\right\}$. If $m < d$, then $m = \left\lfloor \frac{t}{\tau} \right\rfloor \geq \frac{t}{2\tau}$, and

$$\sqrt{\omega(\omega+1)} = \sqrt{\left(\frac{d}{m} - 1\right)\frac{d}{m}} \leq \frac{d}{m} \leq \frac{2d\tau}{t}.$$

If $m = d$, then $\omega = 0$, so the same bound holds. Similarly, if $\ell < d$, then $\ell = \left\lfloor \frac{t}{\kappa} \right\rfloor \geq \frac{t}{2\kappa}$, and

$$\sqrt{\omega_{\text{s}}(\omega_{\text{s}}+1)} = \sqrt{\left(\frac{d}{\ell} - 1\right)\frac{d}{\ell}} \leq \frac{d}{\ell} \leq \frac{2d\kappa}{t}.$$

If $\ell = d$, then $\omega_{\text{s}} = 0$, so this bound also holds. Finally, if $m < d$ and $\ell < d$, then

$$\omega(\omega+1)\omega_{\text{s}} = \left(\frac{d}{m} - 1\right)\frac{d}{m}\left(\frac{d}{\ell} - 1\right) \leq \frac{d^3}{m^2\ell} \leq \frac{8d^3\tau^2\kappa}{t^3}.$$

Otherwise, this product is zero, and the bound remains valid. Therefore, our choice of b, m, and ℓ results in

$$\mathbb{E}\left[T_{\text{time}}\right] = \mathcal{O}\left(\frac{t\Delta}{\varepsilon}\left(\frac{2d\kappa L_A}{t} + \frac{2d\kappa L_B}{t\sqrt{n}} + \left(\frac{2d\tau}{t\sqrt{n}} + \frac{1}{\eta}\right)L_{\max}\right) + \frac{t}{\eta} + d\tau\right)$$

$$= \mathcal{O}\left(\frac{t\Delta}{\varepsilon}\left(\frac{d\kappa L_A}{t} + \frac{d\kappa L_B}{t\sqrt{n}} + \frac{d\tau L_{\max}}{t\sqrt{n}}\right) + \frac{t}{\eta}\left(\frac{\Delta L_{\max}}{\varepsilon} + 1\right) + d\tau\right)$$
$$= \mathcal{O}\left(\frac{\Delta}{\varepsilon}\left(d\kappa L_A + \frac{d\kappa L_B}{\sqrt{n}} + \frac{d\tau L_{\max}}{\sqrt{n}}\right) + \frac{t}{\eta}\left(\frac{\Delta L_{\max}}{\varepsilon} + 1\right) + d\tau\right).$$

By Lemma D.7, we have $\|\nabla f(x_0)\|^2 \leq 2L\Delta$. Assuming that x_0 is not already an ε-stationary point (i.e., $\|\nabla f(x_0)\|^2 \geq \varepsilon$) and noting that $L \leq L_{\max}$, we obtain $\varepsilon \leq 2L_{\max}\Delta$, which implies $\frac{\Delta L_{\max}}{\varepsilon} \geq \frac{1}{2}$. Consequently, $1 \leq \frac{2\Delta L_{\max}}{\varepsilon}$, and therefore $\frac{\Delta L_{\max}}{\varepsilon} + 1 \leq \frac{3\Delta L_{\max}}{\varepsilon}$. Then, we get

$$\mathbb{E}\left[T_{\text{time}}\right] = \mathcal{O}\left(\frac{\Delta}{\varepsilon}\left(d\kappa L_A + \frac{d\kappa L_B}{\sqrt{n}} + \frac{d\tau L_{\max}}{\sqrt{n}}\right) + \frac{t}{\eta} \times \frac{\Delta L_{\max}}{\varepsilon} + d\tau\right). \tag{88}$$

Consider $\frac{t}{\eta}$ separately. Using $b \geq \frac{t}{2h}$ and the compressor bounds above, we obtain

$$\frac{t}{\eta} = t \times \mathcal{O}\left(\sqrt{\frac{\omega(\omega+1)\sigma^2}{n\varepsilon b}} + \frac{\sigma^2}{n\varepsilon b} + \left(\frac{\omega(\omega+1)\omega_{\text{s}}}{n}\right)^{\frac{1}{3}} + 1\right)$$
$$= t \times \mathcal{O}\left(\frac{2d\tau}{t}\sqrt{\frac{\sigma^2}{n\varepsilon}\frac{2h}{t}} + \frac{2\sigma^2 h}{n\varepsilon t} + \left(\frac{8d^3\tau^2\kappa}{nt^3}\right)^{\frac{1}{3}} + 1\right)$$
$$= \mathcal{O}\left(d\tau\sqrt{\frac{\sigma^2 h}{n\varepsilon t}} + \frac{\sigma^2 h}{n\varepsilon} + \left(\frac{d^3\tau^2\kappa}{n}\right)^{\frac{1}{3}} + t\right).$$

Next, we balance the terms $d\tau\sqrt{\frac{\sigma^2 h}{n\varepsilon t}}$ and t:

$$d\tau\sqrt{\frac{h\sigma^2}{n\varepsilon t}} = t \iff t^{\frac{3}{2}} = d\tau\sqrt{\frac{h\sigma^2}{n\varepsilon}} \iff t = \sqrt[3]{\frac{d^2\tau^2 h\sigma^2}{n\varepsilon}}.$$

Thus, choosing $t = \max\left\{h, \tau, \kappa, \sqrt[3]{\frac{d^2\tau^2 h\sigma^2}{n\varepsilon}}\right\}$ we ensure

$$\frac{t}{\eta} = \mathcal{O}\left(\frac{\sigma^2 h}{n\varepsilon} + \left(\frac{d^3\tau^2\kappa}{n}\right)^{\frac{1}{3}} + \max\left\{h, \tau, \kappa, \sqrt[3]{\frac{d^2\tau^2 h\sigma^2}{n\varepsilon}}\right\}\right). \tag{89}$$

Substituting (89) into (88) we bound $\mathbb{E}\left[T_{\text{time}}\right]$ as

$$\frac{\Delta}{\varepsilon} \times \mathcal{O}\left(d\kappa L_A + \frac{d\kappa L_B}{\sqrt{n}} + \left(\frac{d\tau}{\sqrt{n}} + \frac{\sigma^2 h}{n\varepsilon} + \left(\frac{d^3\tau^2\kappa}{n}\right)^{\frac{1}{3}} + \max\left\{h, \tau, \kappa, \sqrt[3]{\frac{d^2\tau^2 h\sigma^2}{n\varepsilon}}\right\}\right) L_{\max}\right)$$
$$+ \mathcal{O}(d\tau).$$

□

G Numerical Estimation of L_A from Assumption 1.6

The parameter L_A plays a crucial role in the time complexity of the methods. Due to Propositions 1.8 and 1.9, we know that it is bounded and is small when the Hessian "does not change too much". In this section, we provide empirical evidence that L_A is small. The code was prepared in Python 3 and executed on a machine with 52 CPUs (Intel(R) Xeon(R) Gold 6278C @ 2.60GHz).

We consider a setup where we train a small convolutional neural network on a two-class classification problem and measure how much the Hessian changes along the optimization trajectory. We use the *MNIST* dataset (LeCun et al., 2010) restricted to two classes. In our experiments, we take the classes 0 and 1, resize all images to 8×8, use 512 random training samples, and use 128 random test samples for estimating the Hessian. The considered neural network consists of one convolutional layer with 4 channels, the *softplus* activation, one hidden linear layer of width 8, another *softplus*

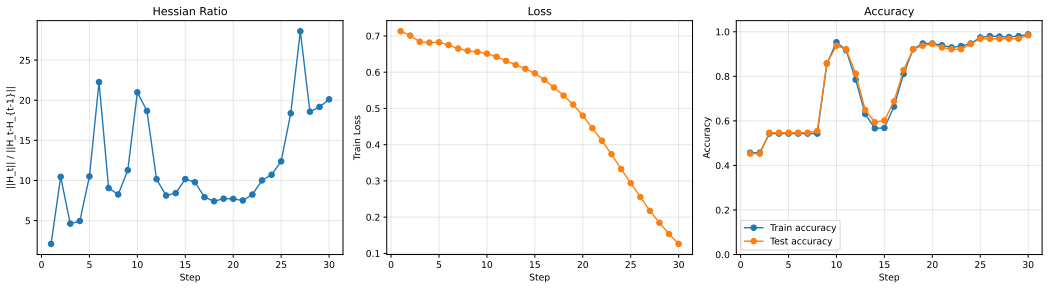


Figure 1: Hessian ratio, loss, and accuracy during training for step size $= 0.01$.

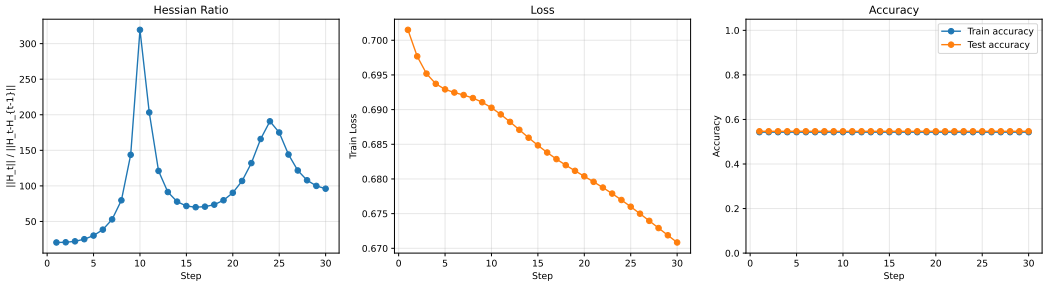


Figure 2: Hessian ratio, loss, and accuracy during training for step size $= 0.001$.

activation, and a final linear layer with two outputs. We use the cross-entropy loss and train the model with Adam (Kingma and Ba, 2015) for 30 iterations.

For each pair of consecutive checkpoints x^{k-1} and x^k, we compute the Hessian of the loss on the test set and report

$$\frac{\|\nabla^2 f(x^k)\|}{\|\nabla^2 f(x^k) - \nabla^2 f(x^{k-1})\|} \approx \frac{L}{L_A},$$

which empirically measures the ratio between local L and L_A due to Proposition 1.9 with $z_1 = x^k$, $z_2 = x^{k-1}$, and $n = 2$. Large values of this ratio indicate that the Hessian changes slowly relative to its norm, which supports the regime where L_A is small. In practice, we estimate L_A locally. However, estimating $\left\|\nabla^2 f(z_1) - \frac{1}{n}\sum_{j=1}^n \nabla^2 f(z_j)\right\|$ is infeasible, even locally, for all $z_1, \ldots, z_n$. Thus, in this section, we assume that the Hessians at the iterates $\{x^k\}_{k\geq 0}$ reflect the correct local geometry.

In Figure 1, we plot the Hessian ratio, loss, and accuracy during training for step size $= 0.01$. We observe that the ratio equals $10 \sim 25$. This provides empirical evidence that the Hessian does not vary significantly during training, and therefore the corresponding local value of L_A can be much smaller than the worst-case smoothness constant L, up to 25 times.

We also repeat the same experiment for step sizes $= 0.001$ and $= 0.1$. With step size $= 0.1$, the method does not converge. In Figure 2, we repeat the experiment for step size $= 0.001$, and observe that the ratio equals $50 \sim 300$, meaning that L_A is smaller than L locally, which is a significant improvement. The reason for this improvement is that the step size is smaller, and therefore the distance between the iterates is smaller. This indicates that our new methods might be especially effective at the last phase of training, when the learning rate is small.

We also consider the experiment from Figure 1 and estimate the ratio using iterates x^k and $x^{k-\delta}$ with $\delta \in \{2, 10\}$ to see how the Hessian distance changes between non-consecutive iterates. In Figures 3 and 4, we observe that the ratio decreases compared to Figure 1. This is expected since the distance between the iterates is larger, and therefore the Hessian can change more.

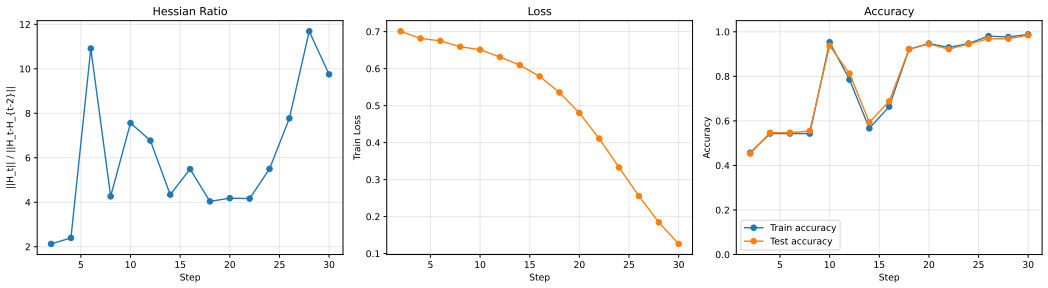


Figure 3: Hessian ratio, loss, and accuracy during training for step size $= 0.01$ and $\delta = 2$.

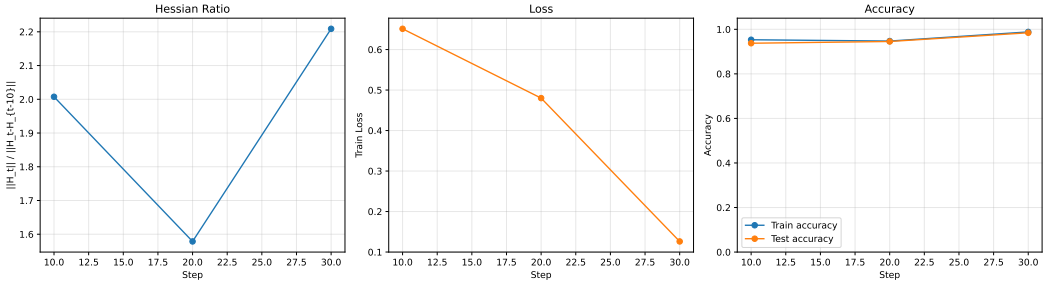


Figure 4: Hessian ratio, loss, and accuracy during training for step size $= 0.01$ and $\delta = 10$.

H Experiments

The experiments were prepared in Python. The distributed environment was emulated on a machine with 52 CPUs (Intel(R) Xeon(R) Gold 6278C @ 2.60GHz).

H.1 Homogeneous Block-Regularized Quadratic

We start our experiments by evaluating the algorithms on a synthetic homogeneous quadratic optimization problem with controlled conditioning. In this setting, we can control different parameters of the problem and make interpretable conclusions. The objective function is defined as

$$f(x) = \frac{1}{2} x^\top \mathbf{A} x, \quad A = \begin{pmatrix} \mathbf{I}_{d/2} & 0 \\ 0 & \lambda \mathbf{I}_{d/2} \end{pmatrix}, \tag{90}$$

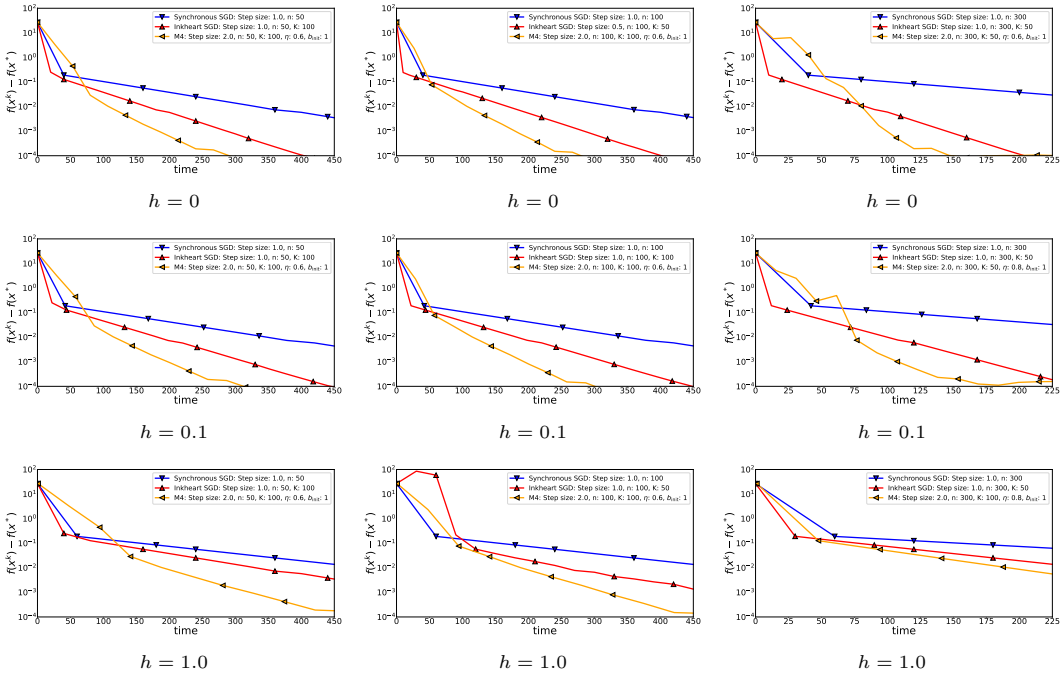
where $d = 300$ is the problem dimension and $\lambda = 0.01$ is a parameter that controls the condition number $\frac{1}{\lambda}$. This construction yields a problem with two blocks of coordinates having distinct curvature: the first $d/2$ coordinates have unit Hessian eigenvalues, while the remaining $d/2$ coordinates are scaled by λ. The unique minimizer is $x^* = 0$, which allows for precise tracking of suboptimality $f(x^k) - f^*$.

Stochastic gradients are obtained by adding controllable Gaussian noise: $\nabla f(x; \xi) = \nabla f(x) + \zeta$, where $\zeta \sim \mathcal{N}(0, \sigma^2)$.

We compare the proposed methods against Synchronous SGD, which is optimal in the centralized distributed optimization with the large noise regime (Tyurin, 2026). In this setup, each worker computes a single stochastic gradient per iteration. Both proposed methods employ Rand$K$ compression for server-worker and worker-server communication, where we tune K from the set $\{1, 3, 10, 30, 50, 100, 200, 300\}$. For all three methods, we tune the step size γ over the range $\{2^{-10}, 2^{-9}, \ldots, 2^3\}$ and report the best performance. In this experiment, we also tune the parameter η from M4 from the set $\{0.1, 0.2, \ldots, 1\}$.

We consider a different number of workers $n \in \{50, 100, 300\}$ under three noise levels: $\sigma = 0.001$ (Figure 5), $\sigma = 0.01$ (Figure 6), and $\sigma = 0.1$ (Figure 7). We observe that the performance of Inkheart SGD and M4 improves as the number of workers increases, which is expected due to the theory. This is especially notable under low computational costs ($h = 0$ and $h = 0.1$), when communication costs

dominate over computational costs, particularly under small noise ($\sigma = 0.001$). In the high noise level $\sigma = 0.1$ (Figure 7), as expected, the gap between the methods decreases.

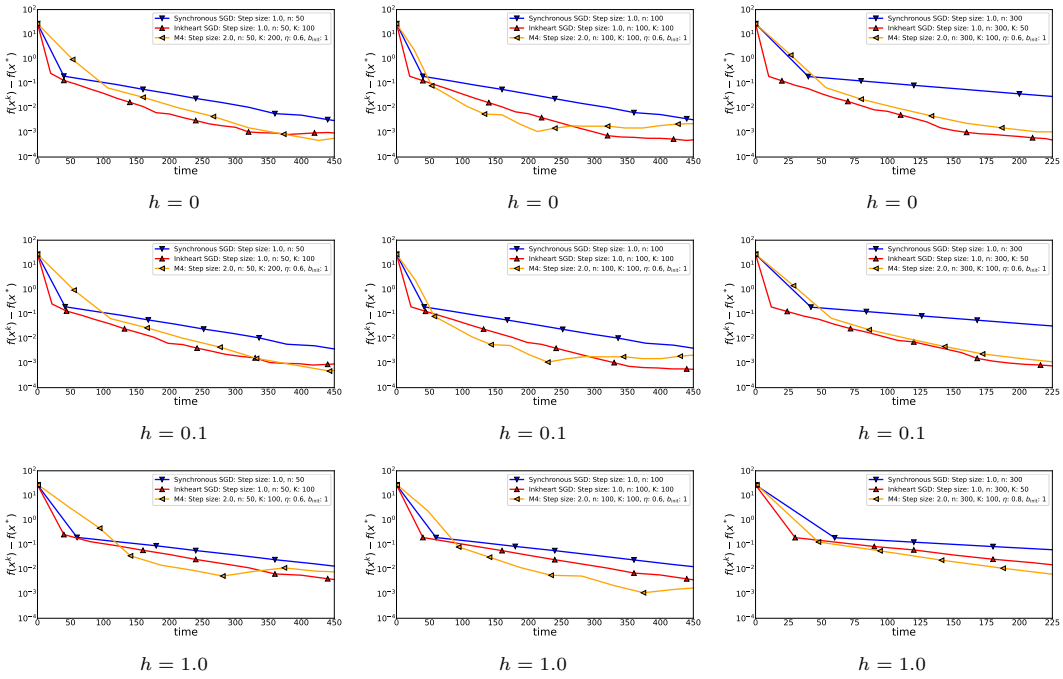


$h = 0$ $h = 0$ $h = 0$

$h = 0.1$ $h = 0.1$ $h = 0.1$

$h = 1.0$ $h = 1.0$ $h = 1.0$

Figure 5: Convergence under low gradient noise ($\sigma = 0.001$). Fixed parameters: $d = 300$, $\kappa = 1/d$, $\tau = 1/d$. Rows vary the gradient computation time h; columns correspond to the number of workers $n \in \{50, 100, 300\}$.

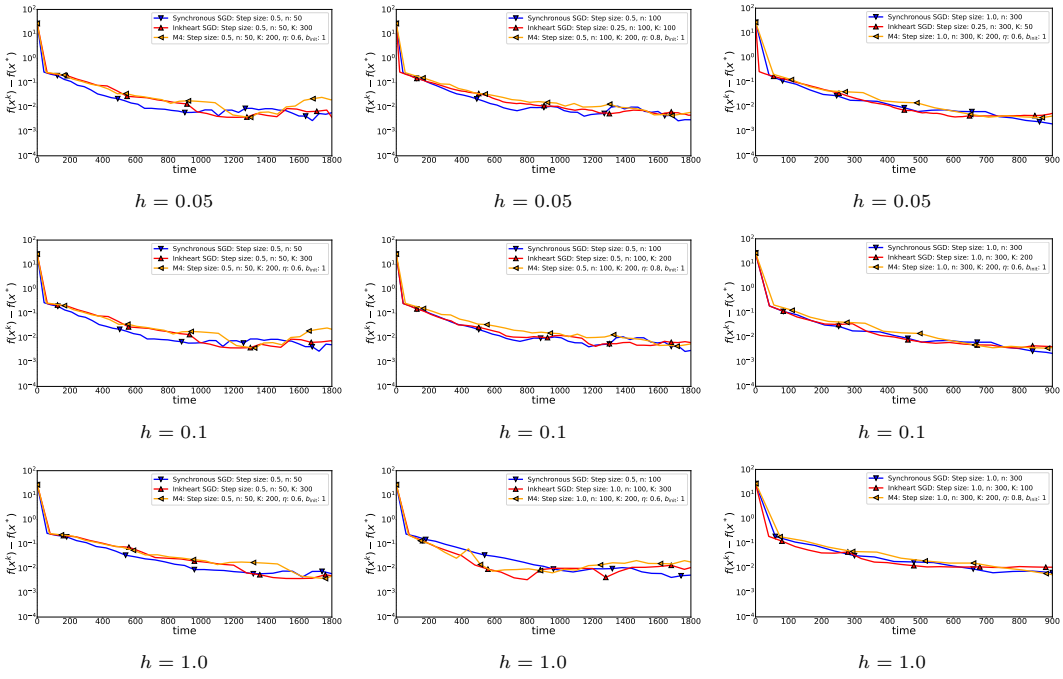


$h = 0$ $h = 0$ $h = 0$

$h = 0.1$ $h = 0.1$ $h = 0.1$

$h = 1.0$ $h = 1.0$ $h = 1.0$

Figure 6: Convergence under medium gradient noise ($\sigma = 0.01$). Fixed parameters: $d = 300$, $\kappa = 1/d$, $\tau = 1/d$. Rows vary the gradient computation time h; columns correspond to the number of workers $n \in \{50, 100, 300\}$.



$h = 0.05$ $h = 0.05$ $h = 0.05$

$h = 0.1$ $h = 0.1$ $h = 0.1$

$h = 1.0$ $h = 1.0$ $h = 1.0$

Figure 7: Convergence under high gradient noise ($\sigma = 0.1$). Fixed parameters: $d = 300$, $\kappa = 1/d$, $\tau = 1/d$. Rows vary the gradient computation time h; columns correspond to the number of workers $n \in \{50, 100, 300\}$.

H.2 Heterogeneous Quadratic Task

Here we consider a heterogeneous quadratic problem. The objective function is almost the same as in (90), but we additionally scale $\mathbf{A}$ to achieve heterogeneity.

$$f_i(x) = \frac{1}{2} x^\top (\xi_i \mathbf{A})\, x, \quad \text{where } \xi_i \sim \mathcal{N}(1, \ell^2) \text{ truncated to } [0.1, 2]. \tag{91}$$

We consider this problem with $\ell = 0.3$ (Figures 8, 10) and $\ell = 0.5$ (Figures 9, 11). As in the previous section, we vary the noise in the stochastic gradients ($\sigma = 0.01, \sigma = 0.001$). We tune hyperparameters K, η, and γ using the same grid as before. We observe that the methods exhibit behavior consistent with the previous experiments.

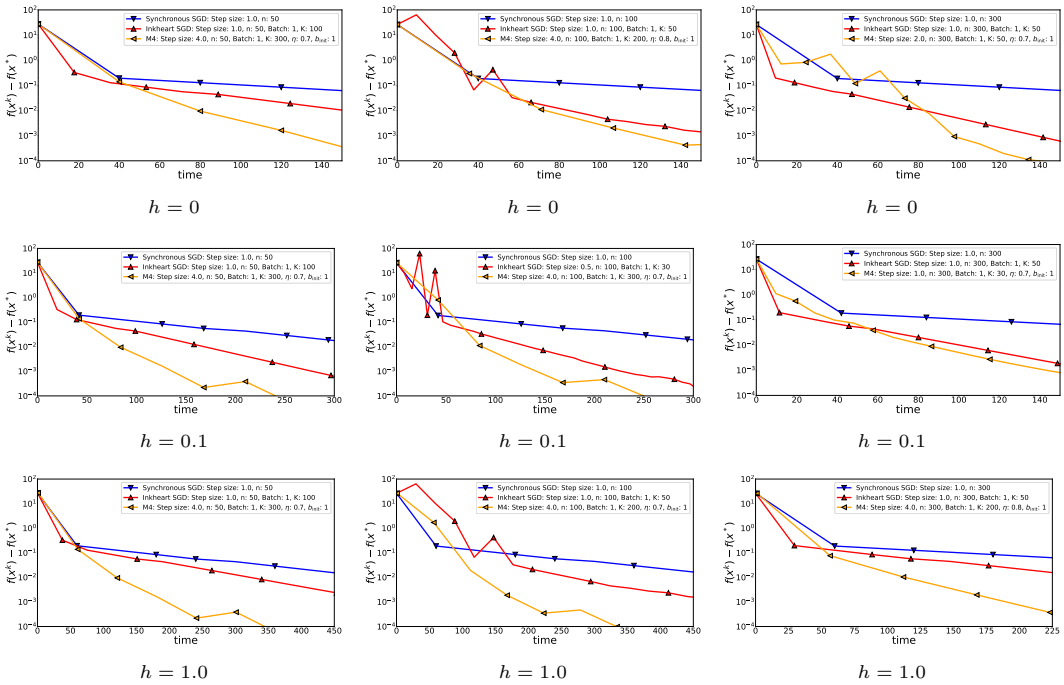


Figure 8: Convergence for medium heterogeneity ($\ell = 0.3$) and low noise $\sigma = 0.001$. Fixed parameters: $d = 300$, $\kappa = 1/d$, $\tau = 1/d$. Rows vary the gradient computation time h; columns correspond to the number of workers $n \in \{50, 100, 300\}$.

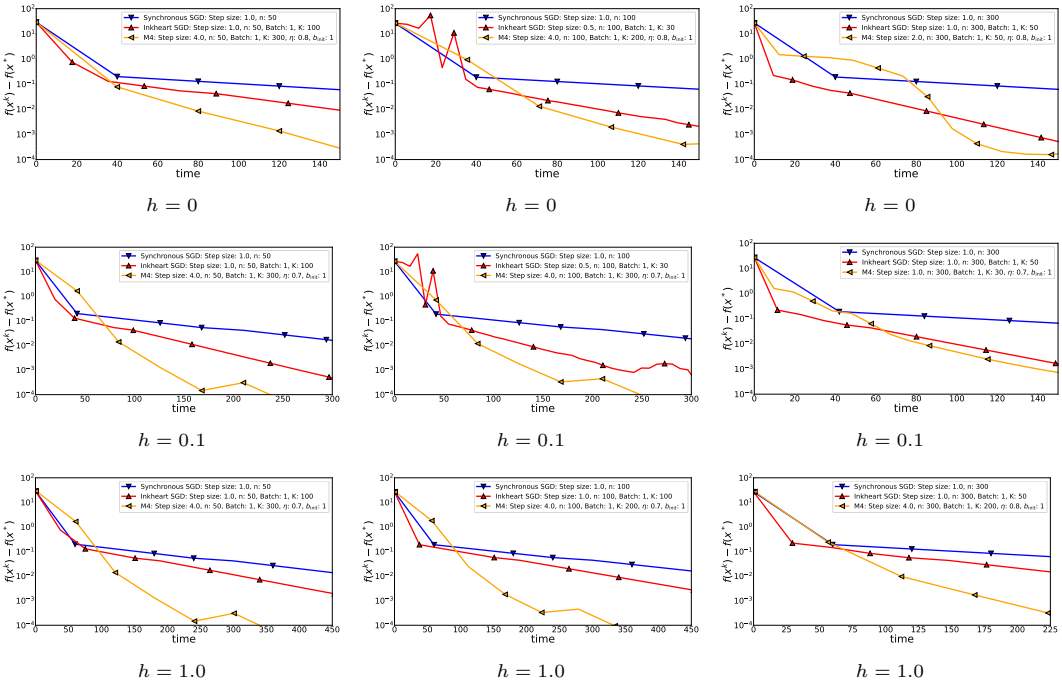


Figure 9: Convergence for high heterogeneity ($\ell = 0.5$) and low noise $\sigma = 0.001$. Fixed parameters: $d = 300$, $\kappa = 1/d$, $\tau = 1/d$. Rows vary the gradient computation time h; columns correspond to the number of workers $n \in \{50, 100, 300\}$.

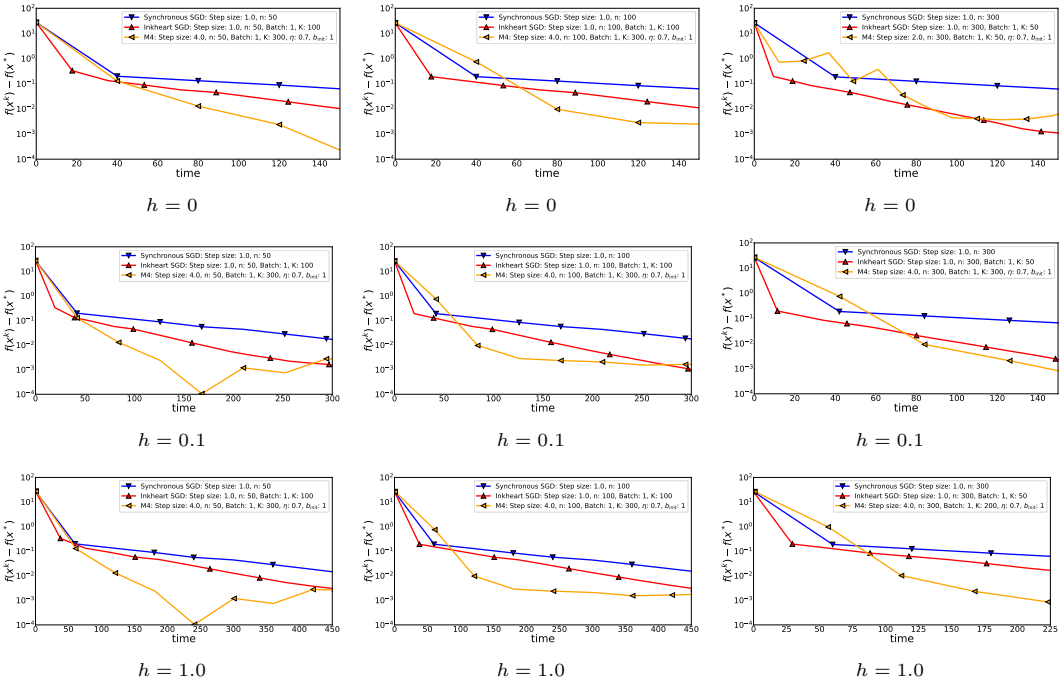
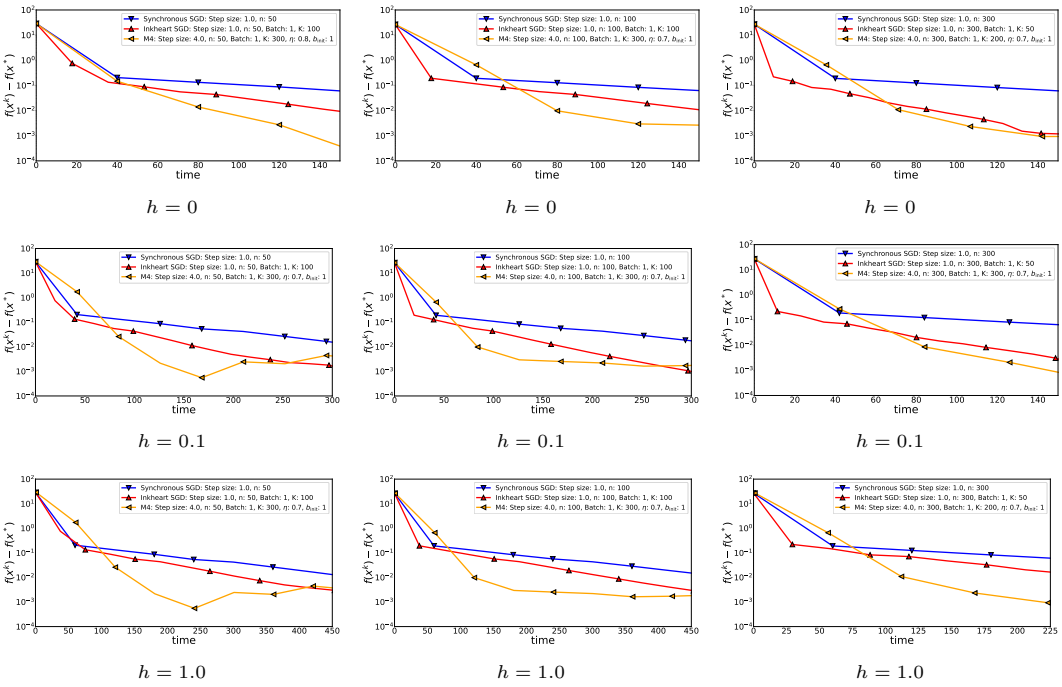


Figure 10: Convergence for medium heterogeneity ($\ell = 0.3$) and medium noise $\sigma = 0.01$. Fixed parameters: $d = 300$, $\kappa = 1/d$, $\tau = 1/d$. Rows vary the gradient computation time h; columns correspond to the number of workers $n \in \{50, 100, 300\}$.



$h = 0$ $h = 0$ $h = 0$

$h = 0.1$ $h = 0.1$ $h = 0.1$

$h = 1.0$ $h = 1.0$ $h = 1.0$

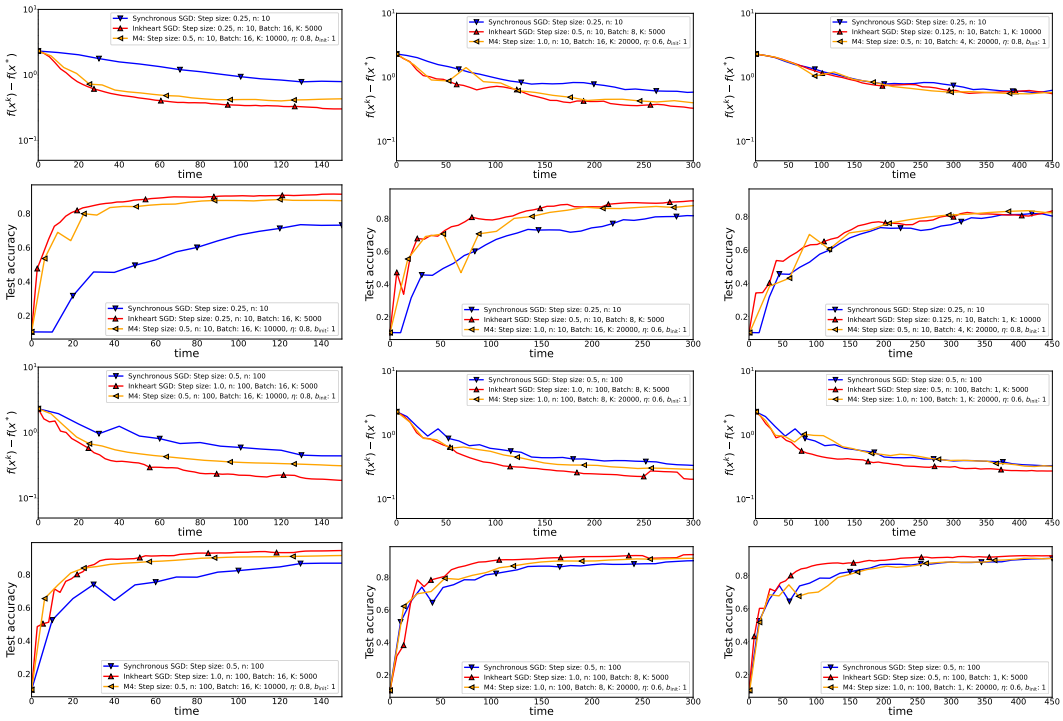
Figure 11: Convergence for high heterogeneity ($\ell = 0.5$) and medium noise $\sigma = 0.01$. Fixed parameters: $d = 300$, $\kappa = 1/d$, $\tau = 1/d$. Rows vary the gradient computation time h; columns correspond to the number of workers $n \in \{50, 100, 300\}$.

H.3 Homogeneous Small-Scale Machine Learning Task

In this section, we train a two-layer neural network (NN) with the architecture $\text{Linear}(\text{input_dim}, 32) \to \text{ReLU} \to \text{Linear}(32, \text{num_classes})$, optimized with the logistic loss on the *MNIST* dataset (LeCun et al., 2010). This setup allows us to compare methods on tasks where workers compute stochastic gradients via uniform sampling, with each worker having access to the full dataset. The total number of network parameters is $d = 25\,450$.

We tune the step size γ over the same range as in the previous sections. We tune the momentum parameter in M4 over $\{0.1, 0.2, \ldots, 1\}$ and take the initial batch size $b_{\text{init}} = 1$. The optimal compression parameter K and the batch size for M4 and Inkheart SGD are selected from the sets $\{100, 1000, 2500, 5000, 7500, \ldots, 25\,000\}$ and $\{1, 4, 8, 16\}$, respectively.

We vary the number of workers $n \in \{10, 100\}$ and the per-sample computation time $h \in \{0, 0.1, 1.0\}$. Figure 12 presents the training loss and accuracy curves. When per-sample computation is cheap, larger batches and aggressive compression enable M4 and Inkheart SGD to outperform Synchronous SGD. Under high computation cost, lighter compression and smaller batches yield better performance.



$h = 0$ $h = 0.1$ $h = 1.0$

Figure 12: Training loss and accuracy on homogeneous MNIST. Columns correspond to per-sample computation time $h \in \{0, 0.1, 1.0\}$. Rows (top to bottom): $n = 10$ (loss), $n = 10$ (accuracy), $n = 100$ (loss), $n = 100$ (accuracy). Fixed parameters: $d = 25\,450$, $\kappa = 1/d$, $\tau = 1/d$.

H.4 Heterogeneous Small-Scale Machine Learning Task

In this section we conduct the same experiment except that *MNIST* is randomly split between workers. The results are presented in Figure 13. We observe that the performance of M4 and Inkheart SGD improves as the number of workers increases. We can see that Inkheart SGD and M4 converge much faster in different computation regimes (different values of h) and for different numbers of workers n, supporting our theoretical results.

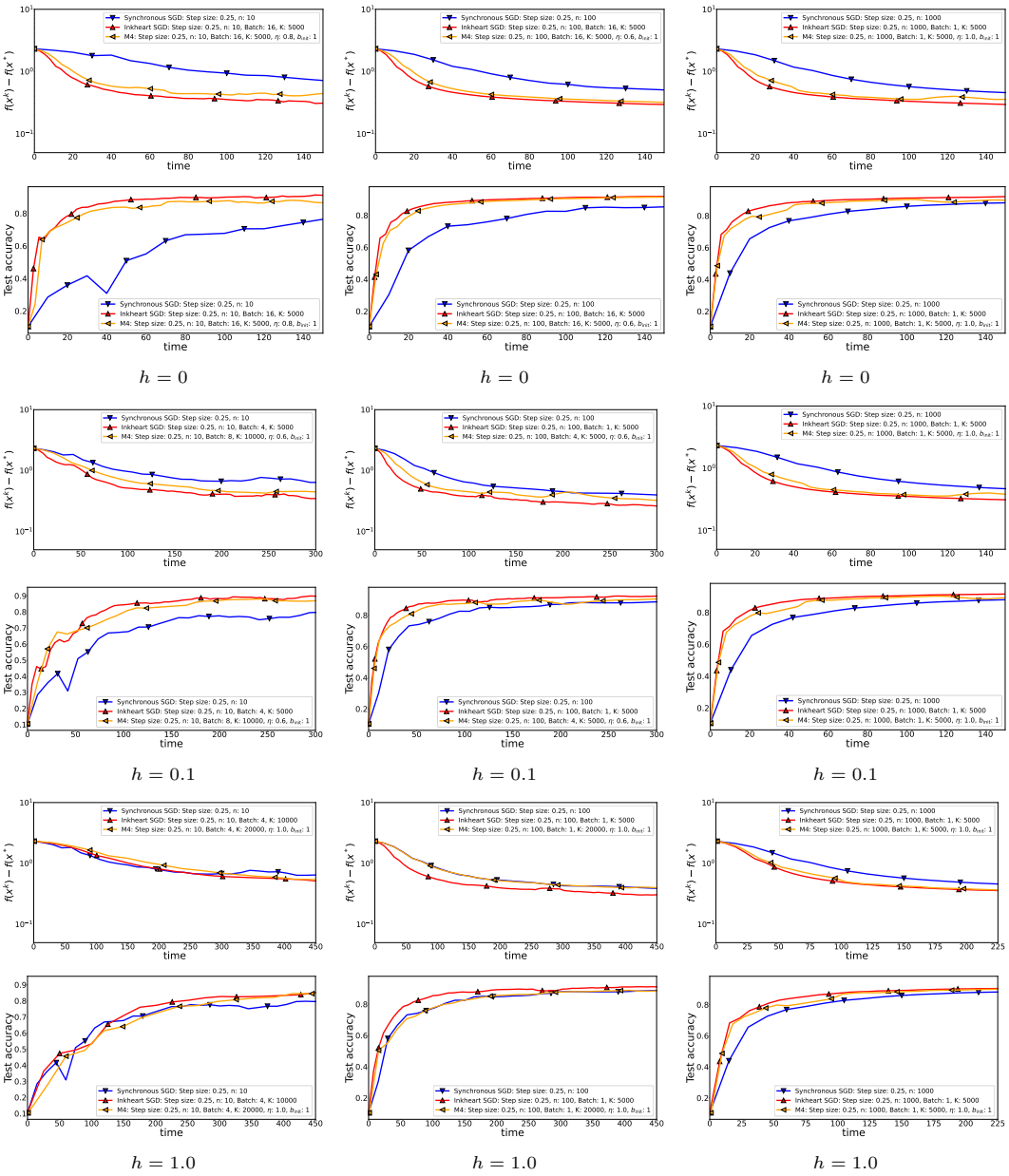


Figure 13: Training loss and accuracy on MNIST. Fixed parameters: $d = 25\,450$, $\kappa = 1/d$, $\tau = 1/d$. Columns vary the number of workers $n \in \{10, 100, 1000\}$; rows alternate between loss and accuracy for the per-sample computation time $h \in \{0, 0.1, 1.0\}$.

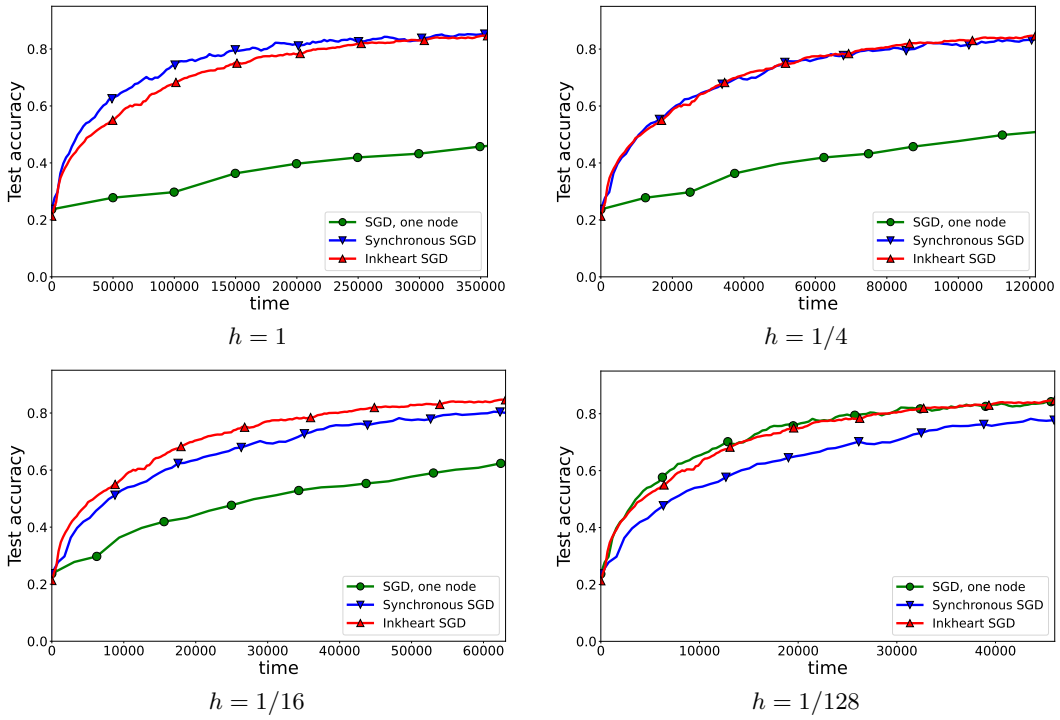


Figure 14: Test accuracy versus simulated execution time for ResNet-18 trained on CIFAR-10. A batch of size 128 is split among the workers. Inkheart SGD and Synchronous SGD use 32 workers, and all 32 workers have access to the full training dataset, with 4 samples per worker. One-node SGD uses a single worker. For Inkheart SGD, $K/d = 0.2$, where K is the compression parameter of Rand$K$. All methods use a constant step size of 0.005. The time required to compute a single-sample stochastic gradient is h seconds, and the communication times satisfy $d\kappa = d\tau = 1$.

H.5 Homogeneous CIFAR-10 experiments with ResNet-18

Additionally, we evaluate the methods on CIFAR-10 (Krizhevsky et al., 2010) using ResNet-18 (He et al., 2016). The resulting model has $d = 11{,}173{,}962$ trainable parameters. Each worker has access to the full training dataset.

We use 32 workers, a global batch size of 128, and hence a local batch size of 4 per worker. All methods are trained for 200 epochs with the same constant step size of 0.005 and weight decay $5 \cdot 10^{-4}$. For Inkheart SGD, we use the compression ratio $K/d = 0.2$. We set $d\kappa = d\tau = 1$, where d is the number of model parameters, and let h denote the computation time of a stochastic gradient for one sample. Thus, each distributed worker spends $4h$ on computation per iteration, while one-node SGD processes all 128 samples sequentially and spends $128h$.

Figure 14 compares the performance under different computation costs. When gradient computation dominates, as in the $h = 1$ regime, Synchronous SGD performs best, since compression provides little benefit in this setting. When computation and communication costs become comparable, Inkheart SGD becomes preferable due to bidirectional compression. Finally, when communication strongly dominates computation, the benefit of distributing the computation becomes smaller, and one-node SGD becomes competitive since it avoids communication completely. Thus, the three regimes exhibit the expected behavior: Synchronous SGD is preferable in computation-dominated settings, Inkheart SGD is preferable in intermediate regimes, and one-node SGD can be preferable when computation is negligible.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Sections 2, 3, and 4

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Sections 2.1, 3.1, and 4.1

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The assumptions and the proofs are in Section 1.4, 3 and in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section H

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: In the supplementary materials.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section H

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification:

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section H

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have reviewed the code of ethics, and we are confident that our paper is in compliance with it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Our work considers a mathematical problem from machine learning.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work considers a mathematical problem from machine learning.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Section H

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code for the experiments is in the supplementary materials.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.