Markovian Compression: Looking to the Past Helps Accelerate the Future

Andrey Veprikov[1,2,3] **Vladimir Solodkin**[1,2,3] **Mikhail Rudakov**[2,3,4] **Petr Babkin**[2] **Aleksandr Beznosikov**[1,2,3,4]

[1]Federated Learning Problems Laboratory
[2]Moscow Independent Research Institute of Artificial Intelligence (MIRAI)
[3]Basic Research of Artificial Intelligence Laboratory (BRAIn Lab)
[4]Innopolis University

Abstract

This paper deals with distributed optimization problems that use compressed communication to achieve efficient performance and mitigate communication bottleneck. We propose a family of compression schemes in which operators transform vectors fed to their input according to a Markov chain, i.e. the stochasticity of the compressors depends on previous iterations. The compressors are implemented in the vanilla Quantized Stochastic Gradient Descent (`QSGD`) algorithm, and, to further improve the efficiency and convergence rate, in the momentum accelerated `QSGD`. We provide convergence results for our algorithms with Markovian compressors, the analysis covers non-convex, Polyak-Lojasiewicz, and strongly convex cases. To demonstrate the applicability of our approach to distributed data-parallel optimization problems, we conduct experiments on the CIFAR-10 and GLUE datasets with the Resnet-18 and DeBERTaV3 models. Practical results show the superiority of methods that use our compressor design over existing schemes.

1 INTRODUCTION

The optimization of high-dimensional objectives has become a cornerstone of modern machine learning, driven by the demands of large-scale neural network training Achiam et al. [2023], distributed resource allocation Sun [2025], and high-frequency algorithmic trading Li et al. [2025]. Consequently, a variety of algorithms for single-device optimization, ranging from classical stochastic methods like `SGD` Robbins and Monro [1951] to adaptive solvers such as `Adam` Kingma and Ba [2015] and the recently proposed `Muon` Jordan et al. [2024], Liu et al. [2025], have emerged and undergone rigorous theoretical analysis. However, in modern deep learning there is a clear shift toward larger and more expressive models, which substantially increases the computational and systems burden of training Team et al. [2025]. This trend is visible across multiple domains, including large-scale vision models for image understanding Zhang et al. [2025], transformer-based architectures in natural language processing Matarazzo and Torlone [2025], and reinforcement learning methods for autonomous decision-making Kiran et al. [2021]. As a result, training is often no longer feasible on a single device: both the required compute and the volume of training data typically exceed the memory and storage constraints of one machine, motivating distributed and data-parallel training setups. Consequently, optimization algorithms have been specifically developed for distributed learning Verbraeken et al. [2020], Chen et al. [2021]. Such type of methods utilize a large number of devices, with each device processing distinct data subsets and participating in an effective data exchange mechanism. This collaborative effort facilitates the training of computationally intensive models. Thus, the problem of classical optimization evolves into a distributed optimization form:

$$\min_{x\in\mathbb{R}^d}\Big\{f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)\Big\}, \tag{1}$$

where f_i is a function, located on a device i. Note that this formulation encompasses two concepts. The first is distributed learning, in which data are spread across multiple devices to facilitate the storage of large amounts of data Liang et al. [2024], Amini et al. [2025]. The second is federated learning Konečnỳ et al. [2016], Li et al. [2020], Kairouz et al. [2021], in which data distribution is motivated by the architecture of the system itself, allowing decentralized model training while maintaining data privacy and integrity between different devices.

The downside of this approach is the complexity associated with the transmission of large-scale data, resulting in "communication bottleneck" Gupta et al. [2021]. This bottleneck can significantly impede the efficiency of the system, particularly in scenarios involving extensive data exchange across

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:6953–6998.

a distributed network. Furthermore, the challenge is amplified in environments with limited bandwidth, requiring solutions to mitigate the impact of data transmission delays and ensure seamless data flow.

The primary solution at present is the compression of the transmitted information Bekkerman et al. [2011], Chilimbi et al. [2014], Alistarh et al. [2017]. That is, no entire package of data is sent, but rather a selected subset. The idea involves strategically selecting and compressing the most informative data segments for transmission. Consequently, this results in the significant advantage of reducing the volume of data that must be communicated across the network, thereby alleviating the communication bottleneck.

Motivated by the aforementioned, a number of methods employing compression have been developed and extensively analyzed Mishchenko et al. [2024], Gorbunov et al. [2021a], Richtárik et al. [2021], Beznosikov et al. [2023a]. However, the majority of studies have focused on unbiased compression operators due to their simplicity and amenability to theoretical analysis. These compression techniques, including random sparsification and value rounding Nesterov [2012a], Alistarh et al. [2017], Horvath et al. [2022], do not in any way take into account the information transmitted in previous iterations. We hence highlight a potential research gap regarding the usage of previously transferred data in compression operators and optimization algorithms.

This omission raises the following research questions that we address in our paper:

• *Is it possible to design compression operators that take into account information about what and how we forwarded in previous iterations?*

• *What methods can we integrate this kind of compression operators into? How does it affect the convergence rate of the methods, both in theory and in practice?*

• *Can the methods be made even more efficient, e.g., by using additional momentum acceleration techniques?*

In this paper, we focus on compression-based methods that take into account information that has already been transferred, employing Markovian compression operators.

1.1 OUR CONTRIBUTIONS

• **New type of compression operator.** We introduce a novel type of compressor that utilizes stochasticity transmitted over previous iterations. We refer to this type of compressors as Markovian because the states of these compressors can be viewed as a Markov chain. Moreover, we provide two naive examples of such compression operators (Definitions 5 and 6) that have been shown to be efficient and simple to implement in real-life problems.

• **In-depth analysis of strongly convex and non-convex cases.** Motivated by various applications primarily from machine learning, we study QSDG and Accelerated QSGD with Markovian compression and provide in-depth discussion of the convergence rate (see Section 2.4). Moreover, we provide theoretical analysis in both the strongly convex (see Theorem 5) and the non-convex / PL-condition (see Theorem 3) case of the target function f.

• **Experimental Validation.** We conduct experiments with Markovian compressors in a data-parallel setup for a series of optimization problems and datasets. In more details, we study the `QSGD`, `Accelerated QSGD` and `DIANA` for logistic regression on MNIST and LIBSVM datasets, and `SGD` for image classification with Resnet-18 on CIFAR-10. Moreover, we consider `Adam` for fine-tuning DeBERTaV3-base for GLUE benchmark. In all setups, we observe an acceleration of convergence for methods employing our compressors compared to the baselines.

1.2 RELATED WORK

Compressed communications. The use of compressed communications is a fairly well-known idea in distributed learning Seide et al. [2014]. As soon as the main property of compressed messages is that they are much easier to transfer, it can be reached in different ways, such as by quantizing the entries of the input vector Alistarh et al. [2017], Mayekar and Tyagi [2021], Gandikota et al. [2021], Horvath et al. [2022], or by sparsifying it Richtárik and Takáč [2016], Alistarh et al. [2018], or even by combining these ideas Albasyoni et al. [2020], Beznosikov et al. [2023a]. However, all of the compression operators could be roughly Condat et al. [2022] separated into two large groups: *unbiased* and *biased*.

The first group is much easier to analyze and is therefore more broadly represented in the literature. The basic method with unbiased compression was presented in Alistarh et al. [2017]. Later this algorithms were modified using variance reduction technique with compression of gradient differences Mishchenko et al. [2024], Horváth et al. [2019], Gorbunov et al. [2021a] in order to improve the theoretical convergence guarantees. One can also note the works Gorbunov et al. [2020] and Khaled et al. [2020], where the authors developed a general theory for `SGD`-type methods with unbiased compression.

On the other hand, theory of distributed optimization with biased compressors is more complicated. In particular, biased compression implies the use of error compensation techniques Stich et al. [2018]. Distributed `SGD` with biased compression and linear rate of convergence in a multi-node setting was first introduced in Beznosikov et al. [2023a]. In the meantime, other error compensation techniques are being actively developed, Lin et al. [2022], Richtárik et al. [2021]. The last approach called `EF21` was later studied in

Fatkhullin et al. [2021], Gruntkowska et al. [2023].

Markovian stochasticity. Another recent trend in the literature is to design algorithms that use Markovian stochastic processes instead of $i.i.d.$ random variables in various ways. For instance, Duchi et al. [2012] introduced a version of the `Mirror Descent` algorithm that yields optimal convergence rates for non-smooth and convex problems. Later, Doan et al. [2020a], Dorfman and Levy [2022], Beznosikov et al. [2023b], Solodkin et al. [2025] studied first-order methods in the Markovian noise setting. Alternatively, token algorithms Hendrikx [2022], Ayache et al. [2022] are also a popular area of research in Markovian stochasticity. In particular, Even [2023] obtained optimal rates of convergence, and Sun et al. [2022], Mao et al. [2020], Doan et al. [2020b] looked at the token algorithm from the angle of the Lagrangian duality and from variants of the `ADMM` method. At the same time, there exist the results, e.g., Bresler et al. [2020], which provide a lower bound for the particular finite sum problems in the Markovian setting.

Despite all of the above, to the best of our knowledge, there are currently no works that combine compressed communications and Markovian stochasticity of the compressors.

1.3 TECHNICAL PRELIMINARIES

Notations. We use $\langle x, y\rangle := \sum_{i=1}^{d} x_i y_i$ to denote standard inner product of vectors $x, y \in \mathbb{R}^d$ and $(x \odot y)_i = x_i y_i$ to denote Hadamard product of vectors $x, y \in \mathbb{R}^d$. We introduce l_2-norm of vector $x \in \mathbb{R}^d$ as $\|x\| := \sqrt{\langle x, x\rangle}$. We define $x^* \in \mathbb{R}^d$ as a point, where we reach the minimum in the problem (1). We also denote $f^* > -\infty$ as a global (potentially not unique) minimum of f. We use a standard notation for $(d-1)$-dimensional simplex $\Delta_d := \left\{p \in \mathbb{R}^d \mid p_j \geq 0 \text{ and } \sum_{j=1}^{d} p_j = 1\right\}$ and for a set of natural numbers $\overline{1,n} := \{1, 2, \ldots, n\}$. We denote C_m^k as the binomial coefficient $\binom{m}{k}$.

Throughout the paper, we assume that the objective functions f_i and the function f from (1) satisfy the following assumptions.

Assumption 1 (L_i-smooth)**.** *Every function f_i is L_i-smooth on $\mathbb{R}^d$ with $L_i > 0$, i.e. it is differentiable and there exists a constant $L_i > 0$ such that for all $x, y \in \mathbb{R}^d$ it holds that $\|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq L_i^2 \|x - y\|^2$. We define $L^2 := \frac{1}{n}\sum_{i=1}^{n} L_i^2$.*

Assumption 2 (μ-strongly convex)**.** *The function f is μ-strongly convex on $\mathbb{R}^d$, i.e., it is differentiable and there is a constant $\mu > 0$ such that for all $x, y \in \mathbb{R}^d$ it holds that $(\mu/2)\|x - y\|^2 \leq f(x) - f(y) - \langle \nabla f(y), x - y\rangle$.*

Assumption 3 (PL-condition)**.** *The function f satisfies the PL-condition, i.e., it is differentiable and there is a constant $\mu > 0$ such that for all $x \in \mathbb{R}^d$ it holds that $\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*)$.*

Assumption 4 (Data similarity)**.** *The functions f_i are similar on $\mathbb{R}^d$, i.e., there are constants $\delta, \sigma \geq 0$, such that the following inequality holds for all $x \in \mathbb{R}^d$: $\|\nabla f_i(x) - \nabla f(x)\|^2 \leq \delta^2 \|\nabla f(x)\|^2 + \sigma^2$.*

The last inequality implies that the data stored at each device does not differ significantly. This Assumption is quite standard in the literature Shamir et al. [2014], Arjevani and Shamir [2015], Khaled et al. [2020], Woodworth et al. [2020], Gorbunov et al. [2021b], Beznosikov et al. [2022, 2023b].

Now we introduce important definitions related to the theory of Markov processes.

Definition 1 (Markov chain)**.** *Markov chain with a finite state space $\{\nu_n\}_{n=0}^N$ is a stochastic process $\{X_t\}_{t\geq 0}$, that satisfies Markov property, i.e. $\mathbb{P}\{X_t = \nu_t \mid X_{t-1} = \nu_{t-1}, X_{t-2} = \nu_{t-2}, ..., X_0 = \nu_0\} = \mathbb{P}\{X_t = \nu_t \mid X_{t-1} = \nu_{t-1}\}$.*

Definition 2 (Ergodicity of Markov chain)**.** *Markov chain $\{X_t\}_{t\geq 0}$ with a finite state space $\{\nu_n\}_{n=0}^N$ is referred to be ergodic if for any $n \in \overline{1,N}$ there exists $\lim_{t\to\infty} \mathbb{P}\{X_t = \nu_n \mid X_0 = \nu_0\} = p_n$, where $0 \leq p_n \leq 1$ does not depend on the ν_0. If Markov chain is ergodic, then $\{p_n\}_{n=0}^N \in \Delta_N$ and there exist $0 < \rho < 1, C > 0$, such that $|\mathbb{P}\{X_t = \nu_n \mid X_0 = \nu_0\} - p_n| \leq C\rho^t$.*

Definition 3 (Mixing time of the discrete Markov chain)**.** *We say that $\tau_{mix}(\varepsilon)$ is the mixing time of the ergodic Markov chain $\{X_t\}_{t\geq 0}$ with stationary distribution $\{p_n\}_{n=0}^N$, if $\forall \varepsilon > 0, \forall t \geq \tau_{mix}(\varepsilon) \hookrightarrow \max_{n\in\overline{0,N}}\{\|\mathbb{P}\{X_t = \nu_n \mid X_0 = \nu_0\} - p_n\|\} \leq \varepsilon \cdot p_{\min}$, where $p_{\min} := \min_{n\in\overline{0,N}}\{p_n\}$. From Definition 2, it follows that $\tau_{mix}(\varepsilon) \geq \frac{\log(C/p_{\min}\varepsilon)}{\log(1/\rho)}$.*

These definitions are extremely important for further analysis of the Markovian compressors, which are presented in the next section.

2 MAIN RESULTS

2.1 MARKOVIAN COMPRESSORS

In this section, we introduce Markovian compressors that take into account the information transmitted in previous K iterations. It is assumed that these compressors function within an iterative algorithm aimed at minimizing the problem (1), wherein a discrete variable t denotes time step. Consequently, due to the dependence of the compressors on previous states, they exhibit a reliance on the step t. Let us narrow down the class of compressors to be discussed in this paper.

Definition 4 (Random sparsification)**.** *$Q_t(x)$ is a random sparsification compressor, if it operates on the vector*

$x \in \mathbb{R}^d$ *as* $Q_t(x) = \frac{d}{m} x \odot \mathbb{1}(\nu_t)$, *where* ν_t *is a set of* m *coordinates:* $\nu_t \subseteq \overline{1,d}$.

The classical Rand$m$ operator fits Definition 4, in particular, subsets ν_t for this compressor are generated uniformly at each step t, therefore it is unbiased, i.e., $\mathbb{E}_t[Q_t(x)] = x$ for all t. In this paper, we do not generate ν_t independently, but according to some Markov chain, i.e., compressors start to take into account previous iterations. We formulate this idea as an assumption.

Assumption 5 (Asymptotic unbiasedness of Markovian compressors)**.** *We assume that operator* Q_t *is a random sparsification compressor (Definition 4) and* $\{\nu_t\}_{t\geq 0}$ *are realizations of some ergodic Markov chain with uniform stationary distribution.*

Assumption 5 implies that in the limit as $t \to \infty$, the compressor Q_t is unbiased, i.e., $\mathbb{E}\left[Q_t(x)\right] \to x$ as $t \to \infty$, because the stationary distribution of the Markov chain is uniform. We are now ready to introduce two examples that adhere to Assumption 5. The first compressor is called $\texttt{BanLast}(K, m)$, it prohibits sending coordinates that have been sent at least once in the last K iterations.

Definition 5 ($\texttt{BanLast}(K, m)$ compressor)**.** *Let* $Q_t(x)$ *be a random sparsification compressor (Definition 4). The* $j \in \nu_t$ *are chosen according to the distribution* $p^t \in \Delta_d$ *and* p^t *is given by the formula:*

$$p_j^t = \begin{cases} 0, & \text{if } j \in \bigcup_{s=t-K}^{t-1} \nu_s, \\ \frac{1}{d-Km}, & \text{otherwise.} \end{cases}$$

The $\texttt{BanLast}(K, m)$ compressor exhibits a limitation in its utility due to an application restriction: $d \geq (K+1)m$, since we need at least m coordinates to have a non-zero probability at each step t. In order to avoid these limitations, we introduce a more flexible Markovian compressor $\texttt{KAWASAKI}(K, b, \pi_\Delta, m)$.

Definition 6 ($\texttt{KAWASAKI}(K, b, \pi_\Delta, m)$ compressor)**.** *Let* $Q_t(x)$ *be a random sparsification compressor (Definition 4). The* $j \in \nu_t$ *are chosen according to the distribution* $p^t \in \Delta_d$*, which is given by the formula:*

$$\widetilde{p}_j^{\,t} = \frac{1/d}{b^{\#\text{ of choices } j \text{ for the last } K \text{ iterations}}}, \quad j \in \overline{1,d};$$
$$p^t = \pi_\Delta\left(\widetilde{p}^{\,t}\right),$$

where $b > 1$ *is a forgetting rate and* $\pi_\Delta : \mathbb{R}^d \to \Delta_d$ *is an activation function.*

The $\texttt{KAWASAKI}(K, b, \pi_\Delta, m)$ compressor is now applicable for arbitrary values of $d \geq m$, and K. On top of that, it introduces two additional parameters in comparison with $\texttt{BanLast}(K, m)$, namely b and π_Δ. The parameter b is responsible for the how strongly we penalize a coordinate if it was selected in previous iterations: the larger b is, the less likely we are to select a coordinate in step t if it was selected in steps from $t-K$ to $t-1$. The function π_Δ is required in order to obtain the probability vector p^t from the vector $\widetilde{p}^{\,t}$. The following examples illustrate potential selections for π_Δ:

$$(\pi_\Delta(\widetilde{p}))_j = |\widetilde{p}_j| / \|\widetilde{p}\|_1,$$
$$\pi_\Delta\left(\widetilde{p}\right) = \text{Softmax}\left(\widetilde{p}\right),$$
$$\pi_\Delta\left(\widetilde{p}\right) = \arg\min_{p \in \Delta_d} \{\|\widetilde{p} - p\|^2\}.$$

We now provide an example where using the Markovian compressor $\texttt{BanLast}(K, m)$ speeds up the optimization process by a factor of three compared to the unbiased compressor Rand$m$.

Example 1. Consider the QSGD algorithm (Algorithm 1), which solves the problem (1) in the case $n = 1$, of the form $x^{t+1} = x^t - \gamma Q(\nabla f(x^t))$. Assume that at some step t we observe gradient of the form $(1, 0, ..., 0)^T \in \mathbb{R}^d$. In the QSGD algorithm, we compress the gradient at each step, therefore, we do not always send the first coordinate to the server, i.e. we do not move from the point x^t.

In the case of $m = 0.1 \cdot d$, i.e. we send 10% of all coordinates at each step, if we use the $\texttt{BanLast}(K, m)$ compressor, then the mathematical expectation of the number of steps to leave the point x^t is approximately 3.4 in the case of $K = 7$. For Rand10%, this number is equal to 10, i.e. we speed up the optimization process by a factor of three. For arbitrary values of d and m, the formula for calculating the number of steps to leave the point x^t is provided in Appendix B.

Moreover, in Appendix B, we obtain the general results for an arbitrary value of $\alpha \in (0; 1]$ with $d = \alpha \cdot m$. For each fixed α, we can find the optimal value of $K^*(\alpha)$. It turns out that empirically this dependence is close to a linear one of the form $K^*(\alpha) \approx 0.73 \cdot \alpha$. It is important to highlight that at this point, K is not treated as a hyperparameter, because it can be selected automatically.

We now present a theorem demonstrating that our Markovian compressors from Definitions 5 and 6 satisfy the conditions outlined in Assumption 5.

Theorem 2 (Asymptotic unbiasedness of $\texttt{BanLast}(K, m)$ and $\texttt{KAWASAKI}(K, b, \pi_\Delta, m)$)**.** *Compressors from Definitions 5 and 6 can be described using Markov chains with states* $\{\nu_1, \nu_2, ..., \nu_K\}_{\nu_1,...,\nu_K \in M}$*, where* M *is the set of all subsets of* $\overline{1,d}$ *of size* m*. Moreover,*

- $\texttt{BanLast}(K, m)$ *(Definition 5) is ergodic with a uniform stationary distribution, if* $d > (K+1)m$.
- *If* $d > (2K+1)m$*, then for* $\texttt{BanLast}(K, m)$ *we get*

$$\rho = \sqrt{1 - \left(\frac{C_{d-2Km}^m}{(C_{d-Km}^m)^2}\right)^K} \text{ and } C = \rho^{-2}.$$

- *If for all permutations ϕ of the set $\overline{1,d}$ it holds that $\pi_\Delta(\phi(\widetilde{p})) = \phi(\pi_\Delta(\widetilde{p}))$, then* `KAWASAKI`$(K, b, \pi_\Delta, m)$ *(Definition 6) is ergodic with a uniform stationary distribution.*
- *If* $(\pi_\Delta(\widetilde{p}))_j = |\widetilde{p}_j|/\|\widetilde{p}\|_1$, *then*

$$\rho = 1 - \left[db^K - m(b^K - 1)\right]^{-mK} \text{ and } C = \rho^{-1}.$$

The proof of Theorem 2 is provided in Appendix C. The outcomes of Theorem 2 hold significant importance for the subsequent investigation of algorithms aimed at solving problem (1) employing Markovian compressors. Note that the examples of activation functions π_Δ provided above satisfy the conditions of Theorem 2.

2.2 DISTRIBUTED GRADIENT DESCENT WITH MARKOVIAN COMPRESSORS

In this section, we propose a new algorithm `Markovian QSGD` (Algorithm 1). This algorithm is similar to the vanilla `QSGD` Alistarh et al. [2017], but in line 5 of Algorithm 1 we use Markovian compressor Q_t^i that we introduced in Section 2.1. That is, Q_t^i can be either `BanLast`(K, m) or `KAWASAKI`(K, b, π_Δ, m), or any other Markovian compressor.

Theorem 3 (Convergence of `MQSGD` (Algorithm 1))**.** *Consider Assumptions 1, 4 and 5. Let the problem* (1) *be solved by Algorithm 1.*

- *For any $\varepsilon, \gamma > 0$ and $T > \tau > \tau_{mix}(\varepsilon)$ such that $\gamma \lesssim \frac{m^2}{d^2 L(\delta^2+1)\tau}$ and $\varepsilon \lesssim \frac{m^2}{d^2(\delta^2+1)}$, it holds that*

$$\mathbb{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right] = \mathcal{O}\left(\frac{F_\tau}{\gamma T} + \frac{\gamma L \tau d^2}{m^2}\sigma^2\right),$$

where $\widehat{x}^T$ is chosen uniformly from $\{x^t\}_{t=0}^T$.

- *If f additionally verifies the PL-condition (Assumption 3), then for any $\varepsilon, \gamma > 0$ and $T > \tau > \tau_{mix}(\varepsilon)$ such that $\gamma \lesssim \frac{m^2}{Ld^2\tau(\delta^2+1)}$ and $\varepsilon = \sqrt{\gamma L \tau} \lesssim \frac{m}{d\sqrt{\delta^2+1}}$, it holds that*

$$F_T = \mathcal{O}\left(\left(1 - \frac{\mu\gamma}{12}\right)^{T-\tau} F_\tau + \frac{\gamma d^2 L \tau}{\mu m^2}\sigma^2\right).$$

Here we use the notations $F_t := \mathbb{E}[f(x^t) - f(x^*)]$ and $F_\tau := \mathbb{E}[f(x^\tau) - f(x^*)]$.

The proof of Theorem 3 is provided in Appendix E.3, E.4. If Assumption 4 does not hold, we observe different results, which are provided in the Appendix F.

Usually in convergence evaluations of various methods, expressions with the term of F_0, i.e., something that depends on the initial guess, arise as constants, but in Theorem 3, a term of the form F_τ appears. This can be explained by the fact that, in iterations from $t = 0$ to $t = \tau$, the Markov chain has not been stabilized yet, and the initial state can be taken as $t = \tau$.

Algorithm 1 `Markovian QSGD`

1: **Input:** starting point $x^0 \in \mathbb{R}^d$, step size $\gamma > 0$, number of iterations T
2: **for** $t = 0$ **to** T **do**
3: Broadcast x^t to all workers
4: **for** $i = 1$ **to** n in parallel **do**
5: Set $g_i^t = Q_t^i(\nabla f_i(x^t))$
6: Send g_i^t to the server
7: **end for**
8: Aggregate $g^t = \frac{1}{n}\sum_{i=1}^n g_i^t$
9: Update $x^{t+1} = x^t - \gamma g^t$
10: **end for**

Sketch proof of Theorem 3. Let us write out a descent lemma of the form:

$$\begin{aligned}\mathbb{E}\left[\|x^{t+1} - x^*\|^2\right] &= \mathbb{E}\left[\left\|x^t - x^*\right\|^2\right] \\ &- 2\mathbb{E}\left[\gamma\left\langle g^t, x^t - x^*\right\rangle\right] + \gamma^2\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n Q_t^i(g_i^t)\right\|^2 \\ &- \underbrace{\frac{2\gamma}{n}\sum_{i=1}^n \mathbb{E}\left[\left\langle Q_t^i(g_i^t) - g_i^t, x^t - x^*\right\rangle\right]}_{①},\end{aligned} \tag{2}$$

where $g^t = \nabla f(x^t)$ and $g_i^t = \nabla f_i(x^t)$ denote the global and local gradients at iteration t, respectively, and Q_t^i represents the Markovian compression operator employed by the i-th worker. The expression ① in (2) is zero if Q_t^i are unbiased and independent from iteration t, because $\mathbb{E}\langle Q_t^i(\nabla f(x^t)) - \nabla f_i(x^t), x^t - x^*\rangle = \mathbb{E}\langle \mathbb{E}_t[Q_t^i(\nabla f(x^t)) - \nabla f_i(x^t)], x^t - x^*\rangle = 0$, where $\mathbb{E}_t[\cdot]$ is the conditional expectation at a step t. Therefore, the theory for such compressors is highly developed. In our case, $Q_t^i(x^s)$ are unbiased only if $t - s \to \infty$, which follows from asymptotic unbiasedness of our Markovian compressors obtained from Assumption 5. However, we can use some coarsening rather than unbiasedness when $t - s = \tau$, where $\tau > \tau_{\text{mix}}(\varepsilon)$, using the technique of *stepping back* as follows:

$$\mathbb{E}\left[\left\langle Q_t^i(a^{t-\tau}) - a^{t-\tau}, b^{t-\tau}\right\rangle\right] \le \frac{\varepsilon d}{m}\mathbb{E}\left[\left\|a^{t-\tau}\right\|\left\|b^{t-\tau}\right\|\right].$$

Importantly, we must apply the compressor Q_t at step t to the vector $a^{t-\tau}$ at step $t - \tau$, since if we apply it to the vector a^t at step t, we will not be able to uncover the conditional expectation, since we will have randomness in a^t (see details in Appendix D). As can be seen from (2) we need to apply the last inequality with $a^{t-\tau} = \nabla f_i(x^{t-\tau})$ and $b^{t-\tau} = x^{t-\tau} - x^*$, but in (2) we only obtain expression with variables at step t, therefore, it has to be handled in

some way. In order to resolve this issue we use a straightforward algebra:

$$\begin{aligned}\mathbb{E}\langle Q_t^i(g_i^t) - g_i^t,\, x^t - x^*\rangle &= \mathbb{E}\langle Q_t^i(g_i^t) - g_i^t,\, x^t - x^{t-\tau}\rangle \\ &+ \mathbb{E}\langle Q_t^i(g_i^{t-\tau}) - g_i^{t-\tau},\, x^{t-\tau} - x^*\rangle \\ &- \mathbb{E}\langle Q_t^i(g_i^t - g_i^{t-\tau}) - g_i^t + g_i^{t-\tau},\, x^t - x^{t-\tau}\rangle \\ &+ \mathbb{E}\langle Q_t^i(g_i^t - g_i^{t-\tau}) - g_i^t + g_i^{t-\tau},\, x^t - x^*\rangle.\end{aligned}$$

The first term in the last inequality is solved with the *stepping back*. The other scalar products are solved using the Fenchel-Young inequality. Terms with $\mathbb{E}\|x^t - x^{t-\tau}\|^2$ are evaluated using line 9 of Algorithm 1: $x^t - x^{t-\tau} = -\gamma\sum_{s=t-\tau}^{t-1} g^s$. The terms with $\mathbb{E}\|Q_t^i(\nabla f_i(x^t) - \nabla f_i(x^{t-\tau}))\|^2$ are obtained from the following inequality (see details in Appendix E):

$$\left\|Q_t^i\left(\nabla f(x) - \nabla f(y)\right)\right\|^2 \leq \frac{d^2L^2}{m^2}\left\|x - y\right\|^2,$$

Since the evaluation of $\mathbb{E}\left\|x^{t+1} - x^*\right\|^2$ raises the terms of the form $\mathbb{E}\left\|x^{t-\tau} - x^*\right\|^2$, we have to do a summation of $\mathbb{E}\left\|x^{t+1} - x^*\right\|^2$ from $t = \tau$ to $t = T$. These terms greatly complicate the proof of Theorem 3 compared to unbiased compressors. The results of Theorem 3 can be rewritten as an upper complexity bound on a number of iterations T of the Algorithm 1 by carefully tuning the step size γ.

Corollary 4 (Step tuning for Theorem 3)**.**

- *Under the conditions of Theorem 3 in the non-convex case, choosing*

$$\gamma \lesssim \frac{m}{d\sqrt{L\tau}}\min\Big\{\frac{m}{d(\delta^2+1)\sqrt{L\tau}}; \sqrt{F_\tau/(T\sigma^2)}\Big\},$$

in order to achieve the ϵ-approximate solution (in terms of $\mathbb{E}\left[\left\|\nabla f(x^T)\right\|^2\right] \leq \epsilon^2$), it takes

$$\mathcal{O}\left(\frac{L\tau d^2}{m^2}F_\tau\left(\frac{\delta^2+1}{\epsilon^2} + \frac{\sigma^2}{\epsilon^4}\right)\right)$$

iterations of Algorithm 1.

- *Under the conditions of Theorem 3 in the PL-condition (Assumption 3) case, choosing*

$$\gamma \lesssim \min\Big\{\frac{m^2}{Ld^2\tau(\delta^2+1)}; \frac{1}{\mu T}\log\big(\max\{2; \frac{\mu^2m^2F_\tau T}{d^2L\tau\sigma^2}\}\big)\Big\},$$

in order to achieve the ϵ-approximate solution (in terms of $\mathbb{E}\left[f(x^t) - f(x^*)\right] \leq \epsilon$), it takes

$$\mathcal{O}\left(\frac{d^2L\tau}{m^2\mu}\left((\delta^2+1)\log\left(\frac{1}{\epsilon}\right) + \frac{\sigma^2}{\mu\epsilon}\right)\right)$$

iterations of Algorithm 1.

Algorithm 2 `Accelerated Markovian QSGD`

Input: starting point $x^0 \in \mathbb{R}^d$, step size $\gamma > 0$, momentums θ, η, β, p, number of iterations T
for $t = 0$ **to** T **do**
  Update $x_g^t = \theta x_f^t + (1-\theta)x^t$
  Broadcast x_g^t to all workers
  for $i = 1$ **to** n in parallel **do**
    Set $g_i^t = Q_t^i\left(\nabla f_i(x_g^t)\right)$
    Send g_i^t to the server
  end for
  Aggregate $g^t = \frac{1}{n}\sum\limits_{i=1}^n g_i^t$
  Update $x_f^{t+1} = x_g^t - p\gamma g^t$
  Update $x^{t+1} = \eta x_f^{t+1} + (p-\eta)x_f^t$
  $\quad + (1-p)(1-\beta)x^t + (1-p)\beta x_g^t$
end for

2.3 ACCELERATED METHOD

After giving the convergence result for the vanilla distributed `SGD` with Markovian compression operator, we now move on to the accelerated scheme. Since we do not assume boundedness of the gradient variance, the classical Nesterov acceleration Nesterov [2014] does not produce the expected effect, and therefore an additional momentum has to be introduced Nesterov [2012b], Vaswani et al. [2019]. By applying a multi-step strategy partially similar to Beznosikov et al. [2023b], we obtain our Algorithm 2.

Theorem 5 (Convergence of `AMQSGD` (Algorithm 2))**.** *Consider Assumptions 1, 2, 4. Let the problem* (1) *be solved by Algorithm 2. Then for any $\gamma, \varepsilon > 0$ and $T > \tau > \tau_{mix}(\varepsilon)$ with*

$$\gamma \lesssim \frac{\mu^{1/3}m^{1/2}}{\tau L^{4/3}d^{1/2}}, \quad p \lesssim \frac{m^2}{\tau^2d^2(\delta^2+1)},$$
$$\varepsilon \lesssim \min\left\{\frac{m^{7/4}}{d^{7/4}\tau^{5/4}L(\delta^2+1)}; \frac{m^{15/4}}{d^{15/4}\tau^{13/4}(\delta^2+1)^2}\right\},$$

and $\beta = \sqrt{2p^2\mu\gamma/3}$, $\eta = \sqrt{3/(2\mu\gamma)}$, $\theta = (p\eta^{-1} - 1)/(\beta p\eta^{-1} - 1)$, it holds that

$$F_{T+1} = \mathcal{O}\Bigg(\exp\left[-(T-\tau)\sqrt{\frac{p^2\mu\gamma}{3}}\right]F_\tau + \exp\left[-T\sqrt{\frac{p^2\mu\gamma}{3}}\right]\Delta_\tau + \frac{\gamma}{\mu}\sigma^2\Bigg).$$

Here we use the notations: $F_t := \mathbb{E}[\|x^t - x^*\|^2 + 3/\mu(f(x_f^t) - f(x^*))]$ and $\Delta_\tau \leq \gamma^{1/2}\tau^{-4/3}\mu^{-1/3}\sum_{t=0}^{\tau}\big(\mathbb{E}\left\|\nabla f(x_g^t)\right\|^2 + \mathbb{E}\left\|x^t - x^*\right\|^2 + \mathbb{E}[f(x_f^t) - f(x^*)]\big)$.

The above theorem shows that in the strongly convex case Algorithm 2 with constant step-size can attain sublinear convergence. In terms of dealing with Markovian stochasticity,

its proof follows quite similar ideas as the proof of Theorem 3: here again we use the technique of *stepping back* for mixing time, which allows us to effectively deal with the bias of the gradient estimator. The full proof is provided in Appendix G.3. The results of Theorem 5 can be rewritten as an upper complexity bound on a number of iterations T of the Algorithm 2 by carefully tuning the step size γ.

Corollary 6 (Step tuning for Theorem 5)**.** *Under the conditions of Theorem 5, choosing*

$$\gamma \lesssim \min\Big\{\frac{\mu^{1/3}}{L^{4/3}\tau^{8/3}}; \frac{1}{\mu p^2 T^2}\log\big(\max\{2; \frac{\mu^{2/3}(F_\tau+\Delta_\tau)T}{\tau^{4/3}L^{2/3}\sigma^2}\}\big)\Big\},$$

in order to achieve the ϵ-approximate solution (in terms of $\mathbb{E}\left[\left\|x^T-x^*\right\|^2\right]\leq\epsilon^2$), it takes

$$\mathcal{O}\left(\frac{d^2L^{\frac{2}{3}}\tau^{\frac{4}{3}}}{m^2\mu^{\frac{2}{3}}}\left((\delta^2+1)\log\left(\frac{1}{\epsilon}\right)+\frac{\sigma^2}{\mu\epsilon}\right)\right)$$

iterations of Algorithm 2.

2.4 DISCUSSION

Our Example 1 and the numerical experiments in Section 3 show that the usage of Markovian compressors could lead to a better performance quite well, however, the theoretical guarantees turn out to be poorer than in the unbiased case. In particular, if we use Rand$m$ in the QSGD algorithm, then we observe the following estimates Beznosikov et al. [2023a]: $X_T=\mathcal{O}((1-\mu\gamma)^T X_0+\gamma\frac{d}{m}\frac{\sigma^2}{\mu n})$, where $X_t=\mathbb{E}\left[\left\|x^t-x^*\right\|^2\right]$ and $\gamma\lesssim\frac{1}{L(1+d/mn)}$. Whereas Theorem 3 gives us such estimates: $F_T=\mathcal{O}(\big(1-\frac{\mu\gamma}{12}\big)^T F_\tau+\gamma\frac{d^2}{m^2}\frac{\tau L\sigma^2}{\mu})$, where $F_t:=\mathbb{E}\left[f(x^T)-f(x^*)\right]$ and $\gamma\lesssim\frac{m^2}{Ld^2\tau(\delta^2+1)}$. It is important to note that not only has the theory for Markovian compressors not yet been studied well, but also dealing with the Markovian stochasticity itself implies quite strict limitations. For instance,

• $\mathbf{d/m}$ **vs** $\mathbf{d^2/m^2}$**.** We are forced to uniformly bound the noise of the compressor (linearity in the compression constant is prevented by this) due to the impossibility of using the expectation trick, in contrast to the unbiased case Beznosikov et al. [2023a], where the authors estimated the variance of the compressor noise. The assumption of uniformly bounded noise cannot be rejected by any authors who work with Markovian stochasticity Beznosikov et al. [2023b], Doan et al. [2020a], Dorfman and Levy [2022], Even [2023], therefore, there is no possibility to achieve linearity in the compression rate in our theoretical guaranties, according to the current theoretical advances.

• **Mixing time.** It is imperative to emphasize that it follows from Theorems 3 and 5 that the convergence rate is improved as τ (and, consequently, K) diminishes. In other words, the distribution of the compressor's underlying Markov chain has to converge to a uniform distribution as fast as possible, but empirically one wants the choice of coordinates to depend on previous iterations rather than be random (e.g., for Rand$m$ compressor $\tau=1, K=0$). This causes a logical contradiction: while using a large K will theoretically give poorer convergence, in practice algorithms with non-zero values of K perform better (see Section 3). It is also worth mentioning that when Markovian stochasticity is employed, we can never avoid τ in the theoretical estimates, since it appears in the lower bounds on the convergence rate of the Markovian methods Bresler et al. [2020]. Thus, our Algorithms 1 and 2 have a reasonably good polynomial dependence on mixing time (Theorem 3 shows an optimal estimation in terms of τ), considering the fact there are several works Doan et al. [2020b] whose bounds include terms that are even *exponential* in the mixing time.

• $\boldsymbol{L/\mu}$**.** In spite of the difficulties listed above, we still can observe that the momentums implementation in Algorithm 2 gives an acceleration in terms of L/μ compared to vanilla QSGD (Algorithm 1). In the classical version of accelerated Gradient Descent, one can achieve an acceleration of the form $\sqrt{L/\mu}$ Nesterov [1983], but our analysis allows only to achieve $(L/\mu)^{2/3}$ in Theorem 5. When Markovian stochasticity is employed, it is also possible to achieve estimation of the form $\sqrt{L/\mu}$ Beznosikov et al. [2023b], but it is obtained by using batches with size scaled as 2^j, where j is drawn from a truncated geometric distribution. Unfortunately, this specific batching technique cannot be applied in our paper, as we consider compressors that act as random sparsification (Definition 4), which necessitates that the gradient be compressed only once at each iteration.

• **Variance reduction.** In our paper, we focus on the QSGD method and its accelerated version (Algorithms 1 and 2). However, in modern studies on distributed optimization, techniques of variance reduction are of a great interest (DIANA Mishchenko et al. [2024], MARINA Gorbunov et al. [2021a], DASHA Tyurin and Richtárik [2023]), because these methods converge linearly to the solution of the problem (1), while QSGD converges only to the σ^2-neighborhood of the solution. We implement Markovian compressors (Definitions 5 and 6) in these methods in our experiments, but we do not provide theoretical guarantees for such algorithms since we have just developed a theoretical baseline for the study of Markovian compressors. This represents a promising direction for future research.

Even though it is not entirely clear whether it is possible to achieve significant improvements in the theoretical results, due to the Markovian randomness, for now we could only highlight a significantly better performance of Algorithms 1 and 2 compared to similar algorithms using a vanilla unbiased compressor Rand$m$ (see Section 3).

3 EXPERIMENTS

In order to justify the practical usage of the proposed methods and analyze their behavior, we conduct a series of experiments using Markovian compression on distributed optimization problems. Despite the considerations mentioned in Section 2.4, we observe that Markovian compressors, when used with `MQSGD` and `AMQSGD`, as well as with classical `SGD`, `DIANA` and `Adam` show better results compared to the baselines. Appendix H provides a description of the technical setup, extended experiments with hyperparameters analysis, and an application of Markovian compressors to model-parallel neural network training.

3.1 LOGISTIC REGRESSION

Firstly, we experiment on a classification task using a logistic regression model with L_2 regularization of the form: $\min_{w\in\mathbb{R}^d}\{f(w) = \frac{1}{n}\sum_{i=1}^{n}\log(1+e^{-y_s w^T x_s}) + \lambda\|w\|^2\}$, with a regularization term $\lambda = 0.05$. The dataset is divided among $n = 10$ clients. We use `Mushrooms`, `A9A`, and `W8A` datasets from LibSVM Chang and Lin [2011] and MNIST Deng [2012]. We experiment with `MQSGD`, `AMQSGD`, and `DIANA` optimizers, employing Rand10% as a sparsification compressor. Markovian compressors were utilized independently on each client, with history size K taken according to theory in Appendix B and forgetting rate $b = 50$. Figure 1 shows the convergence of the baselines and Markovian compressors on the `MQSGD` and `AMQSGD` algorithms on MNIST dataset. Note that improved convergence is achieved without compressor's hyperparameters tuning, making it easy to utilize in any practical optimization task. Moreover, the analysis in Appendix B (see Figure 5) shows that our theory suggests optimal values of history size K.

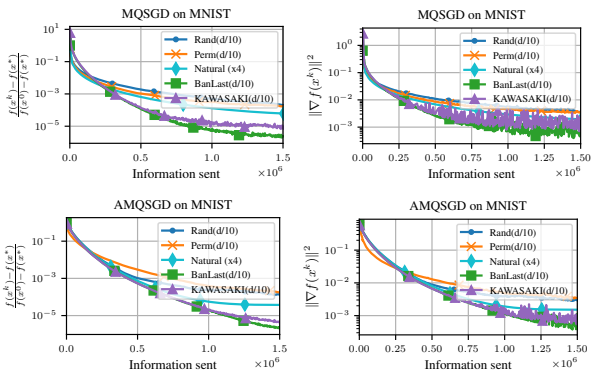

Figure 1: Logistic Regression on MNIST experiments results. Best runs for each method are displayed.

Both Markovian compressors achieve faster convergence than the baseline and more complex compressors: PermK Szlendak et al. [2022] and Natural compression Horvath et al. [2022]. Additionally, we conduct experiments with combination of our Markovian compressors (Definitions 5 and 6) and Natural compression in Appendix H.5.

3.2 COMPUTER VISION

We also apply Markovian compressors in more complex optimization tasks, such as image classification on CIFAR-10 Krizhevsky et al. [2009] dataset with `ResNet-18` convolutional neural network He et al. [2016]. Formally, we solve the optimization problem: $\min_{w\in\mathbb{R}^d}\{f(w) = \frac{1}{n}\sum_{i=1}^{n} l(\sigma(f(x_i, w)), y_i)\}$, where x_i is a training image, y_i is its respective class, $\sigma(\cdot)$ is a Softmax function and $l(\cdot,\cdot)$ is a cross-entropy loss. Dataset is split equally between $n = 5$ clients. Figure 2 depicts the training loss and gradient norm, with the aggregate values shown in Table 1. As in the previous case, the application of the Markovian compressor favors faster convergence and better validation results. Hyperparameters, such as the learning rate and batch size, are fine-tuned.

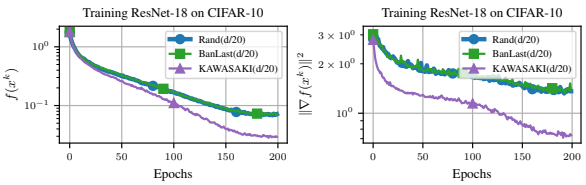

Figure 2: Image classification with ResNet-18 on CIFAR-10 experiments results. Best runs for each method are displayed.

Table 1: Numerical results of training ResNet-18 on CIFAR-10 with different compressors.

	Rand5%	`Banlast`	`KAWASAKI`
Train Loss	0.0743	0.0734	**0.0305**
Gradient Norm	1.403	1.383	**0.745**
Test Accuracy	87.9	88.0	**89.05**

Figure 3 presents comparison with Permutation and Natural compression, which confirm the practical usefulness of Markovian compressors on more complex and non-convex optimization problems. Note that our methods can be applied in combination with sparsification techniques like Natural compression, making our approach even more flexible.

3.3 NATURAL LANGUAGE PROCESSING

Additionally, we perform a series of real-life experiments of fine-tuning DeBERTaV3-base model He et al. [2021] for a subset of the GLUE Wang et al. [2019] benchmark. Figure 4 presents the convergence rate of `Adam` Kingma and Ba [2015] with our Markovian compressors for QNLI and SST2 datasets. A detailed description of this experiment

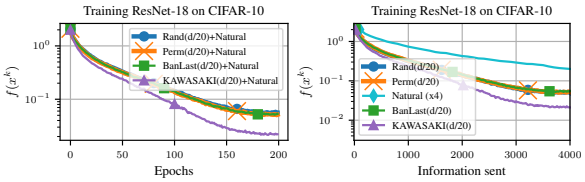

Figure 3: Comparison with other compressors on Resnet-18 training on CIFAR-10 dataset for Rand5% sparsification on $n = 20$ clients. Natural compression factor is 4. Left figure is sequential combination with Natural compression. Right figure is comparison against PermK and Natural compression independently.

is provided in Appendix H.8. Convergence is also notably enhanced in this case in comparison to the baseline.

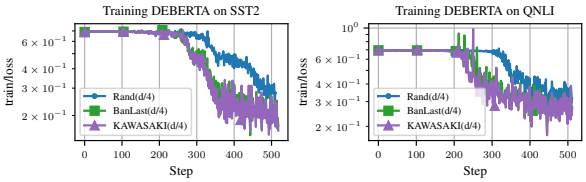

Figure 4: DeBERTaV3 experiments results. Best runs for each method are displayed.

4 CONCLUSION

We studied compression for distributed optimization in the case where the compressor remembers what was sent on earlier iterations. The randomness of such an operator forms a Markov chain, and we collect all operators of this kind into a single class, with `BanLast` and `KAWASAKI` as two concrete members. We embedded the class into `QSGD` and its accelerated version, as well as into `DIANA` and `SGD` with momentum. The convergence guarantees are proved once for an arbitrary compressor of the class under an ergodicity condition, so a new operator only has to satisfy that condition and provide its own constants. The analysis covers the strongly convex, PL, and non-convex regimes. Across logistic regression on MNIST and LIBSVM, image classification with ResNet-18 on CIFAR-10, and fine-tuning of DeBERTaV3 on GLUE, the methods built on our compressors reach a lower loss faster than the random sparsification, permutation, and natural compression baselines.

ACKNOWLEDGMENTS

The work was done in the Laboratory of Federated Learning Problems (Supported by Grant App. No. 2 to Agreement No. 075-03-2024-214).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alyazeed Albasyoni, Mher Safaryan, Laurent Condat, and Peter Richtárik. Optimal gradient compression for distributed and federated learning, 2020.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in neural information processing systems*, 30, 2017.

Dan Alistarh, Torsten Hoefler, Mikael Johansson, Sarit Khirirat, Nikola Konstantinov, and Cédric Renggli. The Convergence of Sparsified Gradient Methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Hadi Amini, Md Jueal Mia, Yasaman Saadati, Ahmed Imteaj, Seyedsina Nabavirazavi, Urmish Thakker, Md Zarif Hossain, Awal Ahmed Fime, and SS Iyengar. Distributed llms and multimodal large language models: A survey on advances, challenges, and future directions. *arXiv preprint arXiv:2503.16585*, 2025.

Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. *Advances in neural information processing systems*, 28, 2015.

Ghadir Ayache, Venkat Dassari, and Salim El Rouayheb. Walk for learning: A random walk approach for federated learning from heterogeneous data, 2022.

Ron Bekkerman, Mikhail Bilenko, and John Langford. *Scaling up machine learning: Parallel and distributed approaches*. Cambridge University Press, 2011.

Aleksandr Beznosikov, Pavel Dvurechenskii, Anastasiia Koloskova, Valentin Samokhin, Sebastian U Stich, and Alexander Gasnikov. Decentralized local stochastic extra-gradient for variational inequalities. *Advances in Neural Information Processing Systems*, 35:38116–38133, 2022.

Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *Journal of Machine Learning Research*, 24(276):1–50, 2023a.

Aleksandr Beznosikov, Sergey Samsonov, Marina Sheshukova, Alexander Gasnikov, Alexey Naumov, and Eric Moulines. First Order Methods with Markovian Noise: from Acceleration to Variational Inequalities. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b.

Song Bian, Dacheng Li, Hongyi Wang, Eric P. Xing, and Shivaram Venkataraman. Does compressing activations help model parallel training?, 2023.

Lukas Biewald. Experiment tracking with weights and biases, 2020. URL `https://www.wandb.com/`. Software available from wandb.com.

Guy Bresler, Prateek Jain, Dheeraj Nagaraj, Praneeth Netrapalli, and Xian Wu. Least Squares Regression with Markovian Data: Fundamental Limits and Algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

Mingzhe Chen, Deniz Gündüz, Kaibin Huang, Walid Saad, Mehdi Bennis, Aneta Vulgarakis Feljan, and H Vincent Poor. Distributed learning in wireless networks: Recent progress and future challenges. *IEEE Journal on Selected Areas in Communications*, 39(12):3579–3605, 2021.

Trishul Chilimbi, Yutaka Suzue, Johnson Apacible, and Karthik Kalyanaraman. Project adam: Building an efficient and scalable deep learning training system. In *11th USENIX symposium on operating systems design and implementation (OSDI 14)*, pages 571–582, 2014.

Laurent Condat, Kai Yi, and Peter Richtárik. EF-BV: A Unified Theory of Error Feedback and Variance Reduction Mechanisms for Biased and Unbiased Compression in Distributed Optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35:30318–30332, 2022.

Michael Diskin, Alexey Bukhtiyarov, Max Ryabinin, Lucile Saulnier, Anton Sinitsin, Dmitry Popov, Dmitry V Pyrkin, Maxim Kashirin, Alexander Borzunov, Albert Villanova del Moral, et al. Distributed deep learning in open collaborations. *Advances in Neural Information Processing Systems*, 34:7879–7897, 2021.

Thinh T. Doan, Lam M. Nguyen, Nhan H. Pham, and Justin Romberg. Convergence rates of accelerated markov gradient descent with applications in reinforcement learning, 2020a.

Thinh T. Doan, Lam M. Nguyen, Nhan H. Pham, and Justin Romberg. Finite-time analysis of stochastic gradient descent under markov randomness, 2020b.

Ron Dorfman and Kfir Y. Levy. Adapting to Mixing Time in Stochastic Optimization with Markovian Data. In *International Conference on Machine Learning (ICML)*, pages 5429–5446. PMLR, 2022.

John C. Duchi, Alekh Agarwal, Mikael Johansson, and Michael I. Jordan. Ergodic Mirror Descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012.

Mathieu Even. Stochastic Gradient Descent under Markovian Sampling Schemes. In *International Conference on Machine Learning (ICML)*, pages 9412–9439. PMLR, 2023.

Ilyas Fatkhullin, Igor Sokolov, Eduard Gorbunov, Zhize Li, and Peter Richtárik. Ef21 with bells & whistles: Practical algorithmic extensions of modern error feedback, 2021.

Venkata Gandikota, Daniel Kane, Raj Kumar Maity, and Arya Mazumdar. vqSGD: Vector Quantized Stochastic Gradient Descent. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2021.

Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A Unified Theory of SGD: Variance Reduction, Sampling, Quantization and Coordinate Descent. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 680–689. PMLR, 2020.

Eduard Gorbunov, Konstantin P Burlachenko, Zhize Li, and Peter Richtárik. Marina: Faster non-convex distributed learning with compression. In *International Conference on Machine Learning*, pages 3788–3798. PMLR, 2021a.

Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. Local sgd: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, pages 3556–3564. PMLR, 2021b.

Kaja Gruntkowska, Alexander Tyurin, and Peter Richtárik. EF21-P and Friends: Improved Theoretical Communication Complexity for Distributed Optimization with Bidirectional Compression. In *International Conference on Machine Learning (ICML)*. PMLR, 2023.

Vipul Gupta, Avishek Ghosh, Michal Derezinski, Rajiv Khanna, Kannan Ramchandran, and Michael Mahoney. Localnewton: Reducing communication bottleneck for distributed learning. *arXiv preprint arXiv:2105.07320*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations (ICLR)*, 2021.

Hadrien Hendrikx. A principled framework for the design and analysis of token algorithms, 2022.

Samuel Horvath, Chen-Yu Ho, Ludovit Horvath, Atal Narayan Sahu, Marco Canini, and Peter Richtarik. Natural Compression for Distributed Deep Learning. In *Mathematical and Scientific Machine Learning (MSML)*, pages 129–141. PMLR, 2022.

Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic distributed learning with gradient quantization and variance reduction, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2022.

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

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

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 4519–4529. PMLR, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2021.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854, 2020.

Yang Li, Zhi Chen, and Steve Yang. Flowhft: Flow policy induced optimal high-frequency trading under diverse market conditions. *arXiv e-prints*, pages arXiv–2505, 2025.

Feng Liang, Zhen Zhang, Haifeng Lu, Victor Leung, Yanyi Guo, and Xiping Hu. Communication-efficient large-scale distributed deep learning: A comprehensive survey. *arXiv preprint arXiv:2404.06114*, 2024.

Chung-Yi Lin, Victoria Kostina, and Babak Hassibi. Differentially Quantized Gradient Methods. *IEEE Transactions on Information Theory*, 68(9):6078–6097, 2022.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for llm training, 2025. URL https://arxiv.org/abs/2502.16982.

Xianghui Mao, Kun Yuan, Yubin Hu, Yuantao Gu, Ali H. Sayed, and Wotao Yin. Walkman: A Communication-Efficient Random-Walk Algorithm for Decentralized Optimization. *IEEE Transactions on Signal Processing*, 68: 2513–2528, 2020.

Andrea Matarazzo and Riccardo Torlone. A survey on large language models with some insights on their capabilities and limitations, 2025. URL https://arxiv.org/abs/2501.04040.

Prathamesh Mayekar and Himanshu Tyagi. RATQ: A Universal Fixed-Length Quantizer for Stochastic Optimization. *IEEE Transactions on Information Theory*, 67(5): 3130–3154, 2021.

Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed Learning with Compressed Gradient Differences. *Optimization Methods and Software*, 2024.

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

Yurii Nesterov. A method of solving a convex programming problem with convergence rate o (1/k** 2). *Doklady Akademii Nauk SSSR*, 269(3):543, 1983.

Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM J. Optim.*, 22:341–362, 2012b. URL https://api.semanticscholar.org/CorpusID:1424102.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014. ISBN 1461346916.

J v Neumann. Proof of the quasi-ergodic hypothesis. *Proceedings of the National Academy of Sciences*, 18(1): 70–82, 1932.

Peter Richtárik and Martin Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156:433–484, 2016.

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. Ef21: A new, simpler, theoretically better, and practically faster error feedback. *Advances in Neural Information Processing Systems*, 34:4384–4396, 2021.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Interspeech*, 2014. URL https://api.semanticscholar.org/CorpusID:2189412.

Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pages 1000–1008. PMLR, 2014.

Vladimir Solodkin, Andrew Veprikov, and Aleksandr Beznosikov. Methods for optimization problems with markovian stochasticity and non-euclidean geometry, 2025. URL https://arxiv.org/abs/2408.01848.

Jaeyong Song, Jinkyu Yim, Jaewon Jung, Hongsun Jang, Hyung-Jin Kim, Youngsok Kim, and Jinho Lee. Optimus-cc: Efficient large nlp model training with 3d parallelism aware communication compression. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pages 560–573, 2023.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. *Advances in Neural Information Processing Systems*, 31, 2018.

Ling Sun. Research on adaptive distributed optimal resource allocation consistency algorithm in wan environment. *Scientific Reports*, 15(1):41245, 2025.

Tao Sun, Dongsheng Li, and Bao Wang. Adaptive random walk gradient descent for decentralized optimization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20790–20809. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/sun22b.html.

Rafał Szlendak, Alexander Tyurin, and Peter Richtárik. Permutation Compressors for Provably Faster Distributed Nonconvex Optimization. In *International Conference on Learning Representations (ICLR)*, 2022.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, Yichen Feng, Kelin Fu, Bofei Gao, Hongcheng Gao, Peizhong Gao, Tong Gao, Xinran Gu, Longyu Guan, Haiqing Guo, Jianhang Guo, Hao Hu, Xiaoru Hao, Tianhong He, Weiran He, Wenyang He, Chao Hong, Yangyang Hu, Zhenxing Hu, Weixiao Huang, Zhiqi Huang, Zihao Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yongsheng Kang, Guokun Lai, Cheng Li, Fang Li, Haoyang Li, Ming Li, Wentao Li, Yanhao Li, Yiwei Li, Zhaowei Li, Zheming Li, Hongzhan Lin, Xiaohan Lin, Zongyu Lin, Chengyin Liu, Chenyu Liu, Hongzhang Liu, Jingyuan Liu, Junqi Liu, Liang Liu, Shaowei Liu, T. Y. Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yibo Liu, Yiping Liu, Yue Liu, Zhengying Liu, Enzhe Lu, Lijun Lu, Shengling Ma, Xinyu Ma, Yingwei Ma, Shaoguang Mao, Jie Mei, Xin Men, Yibo Miao, Siyuan Pan, Yebo Peng, Ruoyu Qin, Bowen Qu, Zeyu Shang, Lidong Shi, Shengyuan Shi, Feifan Song, Jianlin Su, Zhengyuan Su, Xinjie Sun, Flood Sung, Heyi Tang, Jiawen Tao, Qifeng Teng, Chensi Wang, Dinglu Wang, Feng Wang, Haiming Wang, Jianzhou Wang, Jiaxing Wang, Jinhong Wang, Shengjie Wang, Shuyi Wang, Yao Wang, Yejie Wang, Yiqin Wang, Yuxin Wang, Yuzhi Wang, Zhaoji Wang, Zhengtao Wang, Zhexu Wang, Chu Wei, Qianqian Wei, Wenhao Wu, Xingzhe Wu, Yuxin Wu, Chenjun Xiao, Xiaotong Xie, Weimin Xiong, Boyu Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Xiaofei Yang, Ying Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Xingcheng Yao, Wenjie Ye, Zhuorui Ye, Bohong Yin, Longhui Yu, Enming Yuan, Hongbang Yuan, Mengjie Yuan, Haobing Zhan, Dehao Zhang, Hao Zhang, Wanlu Zhang, Xiaobin Zhang, Yangkun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Haotian Zhao, Yikai Zhao,

Huabin Zheng, Shaojie Zheng, Jianren Zhou, Xinyu Zhou, Zaida Zhou, Zhen Zhu, Weiyu Zhuang, and Xinxing Zu. Kimi k2: Open agentic intelligence, 2025. URL https://arxiv.org/abs/2507.20534.

Alexander Tyurin and Peter Richtárik. DASHA: Distributed Nonconvex Optimization with Communication Compression, Optimal Oracle Complexity, and No Client Synchronization. In *International Conference on Learning Representations (ICLR)*, 2023.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and Faster Convergence of SGD for Over-Parameterized Models and an Accelerated Perceptron. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2019.

Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S Rellermeyer. A survey on distributed machine learning. *Acm computing surveys (csur)*, 53(2):1–33, 2020.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations (ICLR)*, 2019.

Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33:6281–6292, 2020.

Zherui Zhang, Rongtao Xu, Jie Zhou, Changwei Wang, Xingtian Pei, Wenhao Xu, Jiguang Zhang, Li Guo, Longxiang Gao, Wenbo Xu, and Shibiao Xu. Image recognition with online lightweight vision transformer: A survey, 2025. URL https://arxiv.org/abs/2505.03113.

Markovian Compression: Looking to the Past Helps Accelerate the Future (Supplementary Material)

Andrey Veprikov[1,2,3] **Vladimir Solodkin**[1,2,3] **Mikhail Rudakov**[2,3,4] **Petr Babkin**[2] **Aleksandr Beznosikov**[1,2,3,4]

[1]Federated Learning Problems Laboratory
[2]Moscow Independent Research Institute of Artificial Intelligence (MIRAI)
[3]Basic Research of Artificial Intelligence Laboratory (BRAIn Lab)
[4]Innopolis University

A AUXILIARY LEMMAS AND FACTS

In this section we list auxiliary facts and our results that we use several times in our proofs.

A.1 CAUCHY–SCHWARZ INEQUALITY

For all $x, y \in \mathbb{R}^d$

$$\langle x, y\rangle \leq \|x\| \, \|y\| \, .$$

A.2 FENCHEL-YOUNG INEQUALITY

For all $x, y \in \mathbb{R}^d$ and $\beta > 0$

$$2 \langle x, y\rangle \leq \beta^{-1}\|x\|^2 + \beta\|y\|^2.$$

B CALCULATIONS FROM EXAMPLE 1

By definition of the mathematical expectation of an integer positive random variable Z, we obtain that $\mathbb{E}[Z] = \sum_{s=1}^{\infty} s \cdot \mathbb{P}\{Z = s\}$. In our problem, Z is the number of an iteration where we first selected the desired coordinate. For Rand$m$ compressor, we have $\mathbb{P}\{Z = s\} = \frac{m}{d} \cdot \left(1 - \frac{m}{d}\right)^{s-1}$. The first term is the probability of picking the desired coordinate at iteration s and the second term is the probability of not picking the desired coordinate at iterations from 1 to $s - 1$. Using this, the mathematical expectation of the number of steps to quit the point x^t for Rand$m$ compressor is equal to

$$\sum_{s=1}^{\infty} s \left(1 - \frac{m}{d}\right)^{s-1} \frac{m}{d} = \frac{d}{m}. \tag{3}$$

Now we calculate the expectation for $\texttt{BanLast}(K, m)$ compressor (Definition 5). If $s > K$, similarly to the Rand$m$ case, we obtain that $\mathbb{P}\{Z = s\} = \frac{m}{d-Km} \left(1 - \frac{m}{d-Km}\right)^{s-1}$, because we cannot choose Km coordinates. If $s \leq K$, then the formula of $\mathbb{P}\{Z = s\}$ becomes a bit more complicated, because the probability of not picking the desired coordinate at iterations from 1 to $s - 1$ is different at each iteration and is equal to $\prod_{h=0}^{s-2} \left(1 - \frac{m}{d-hm}\right)$. If $s = 1$, then this probability is

equal to one. Using this, we can calculate the mathematical expectation of the number of steps to leave the point x^t for $\texttt{BanLast}(K, m)$ compressor:

$$\begin{aligned}
&\sum_{s=1}^{K} \frac{sm}{d-(s-1)m} \prod_{h=0}^{s-2}\left(1-\frac{m}{d-hm}\right) + \sum_{s=K+1}^{\infty} s\left(1-\frac{m}{d-Km}\right)^{s-1}\frac{m}{d-Km} \\
&= \sum_{s=1}^{K} \frac{sm}{d-(s-1)m} \prod_{h=0}^{s-2}\left(1-\frac{m}{d-hm}\right) + \frac{d}{m}\left(1-\frac{m}{d-Km}\right)^{K} \\
&= \sum_{s=1}^{K} \frac{s}{\alpha-(s-1)} \prod_{h=0}^{s-2}\left(1-\frac{1}{\alpha-h}\right) + \alpha\left(1-\frac{1}{\alpha-K}\right)^{K},
\end{aligned} \tag{4}$$

where we used the notation $\alpha = d/m$ to show that (4) depends only on d/m, but not on d and m separately. We can consider (4) as an optimization problem with respect to K. Since K is an integer and the objective function in (4) is complex, we numerically find the optimal K for different α. For the sake of clarity, we show the difference between formulas (3) and (4) on Figure 5(c).

We consider $\alpha \in [5.3, 6.7, 8.3, 10, 11.1, 12.5, 14.3, 16.7, 20]$ and find the optimal K by a complete brute force search – see Figure 5 (a). Then, we perform a linear approximation and obtain the formula $K^*(\alpha) \approx 0.7323\alpha$ – see Figure 5 (b). Since the correlation coefficient between the points and the approximated line is equal to 0.73, we can consider this formula to be accurate enough for practical applications.

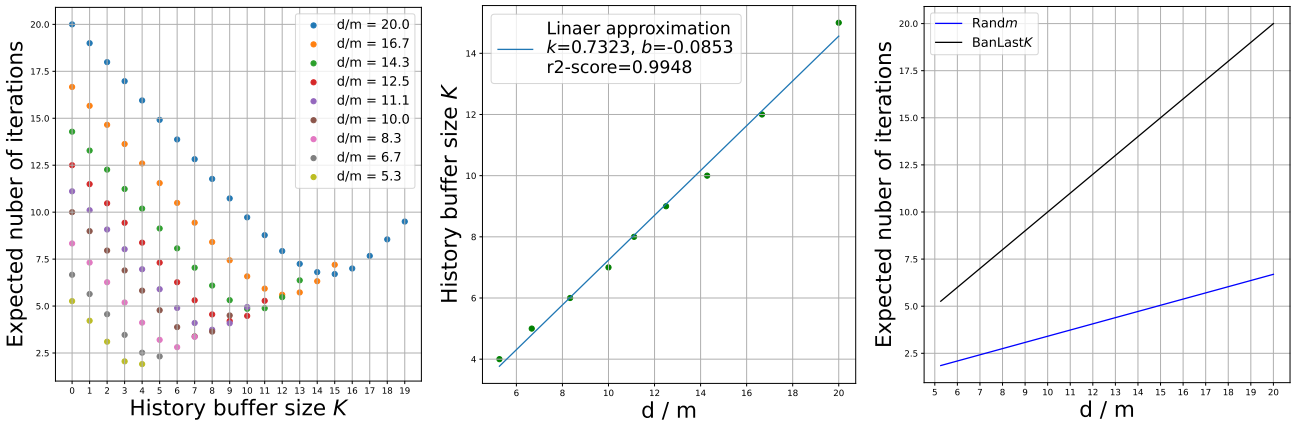

Figure 5: Theoretical estimate on dependence of history buffer size K on parameter $\alpha = d/m$: (a) represents expected number of iterations required to transfer all coordinates to server on history buffer size K for different α, (b) represents scaling of optimal history buffer size K^* on α. (c) represents comparison of expected number of iterations required to transfer all coordinates to server on problems parameter α for Rand$m$ and BanLast$K$.

C PROOF OF THEOREM 2

Lemma 7. *If P is a transition matrix of a finite homogeneous Markov chain, i.e.*

$$P := (p_{ij})_{i,j=1}^{n},$$

where p_{ij} is probability of moving from i to j in one time step. And the matrix P is symmetric, i.e. $P^T = P$, then stationary distribution exists and it is uniformly distributed.

Proof of Lemma 7.. Let us look at uniform distribution

$$\pi := \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right).$$

We can easily obtain that π is a stationary distribution, using symmetry and stochastic property of matrix P:

$$\pi P = \frac{1}{n}\mathbf{1}^T P = \frac{1}{n}(P\mathbf{1})^T = \frac{1}{n}\mathbf{1}^T = \pi.$$

□

Proof of Theorem 2.. We consider states of Markov chain as $s := \{\nu_1, \nu_2, ..., \nu_K\}_{\nu_1,...,\nu_K \in M}$, where M is the set of all subsets of $\overline{1,d}$ of size m. We define $p(s, s', i)$ as the probability to move from state s to state s' for the number of steps i.

• For both compressors $\texttt{BanLast}(K, m)$ (Definition 5) and $\texttt{KAWASAKI}(K, b, \pi_\Delta, m)$ (Definition 6) corresponding Markov chain is finite and indecomposable.

The finiteness of the chain is apparent, as the number of states can be explicitly expressed as $|M| = (C_d^m)^K$. We show that both chains are indecomposable below. Then we deduce that the chain is ergodic based on the Ergodic Theorem Neumann [1932]. Thus, we know that a stationary distribution exists. Than we show that the statinary distribution is uniform over the set of states using Lemma 7.

All that remains is to show that both chains are indecomposable and that transition matrixes for both chaines are symmetric.

We will start with $\texttt{BanLast}(K, m)$. Restriction on K, m and d is $d > (K+1)m$. That makes obvious that any two states are communicated, i.e. for any s, s' there exists way from s to s'. Thus, the Markov chain is indecomposable.

For the compressor probability to move from s to s' in one time step can be explicitly expressed as:

$$p(s, s', 1) = \left(\frac{1}{C_{d-Km}^m}\right)^K,$$

where $C_{d-Km}^m = \frac{(d-Km)!}{m!(d-(K+1)m)!}$ is a binomial coefficient. And all these states are equal in probability. If $d = (K+1)m$, then for s there will be only one set s', such that $p(s, s', 1) > 0$, in this case chain will not be ergodic. If $d > (K+1)m$, then there are more then one state s', for witch $p(s, s', 1) > 0$, therefore chain will be ergodic.

• According to the Ergodic Theorem, $\rho = (1-\delta)^{1/N_0}$ and $C = (1-\delta)^{-1}$, where N_0 is the minimal number of iterations through which is strictly greater then zero and $\delta := \min_{s,s'}\{p(s, s', N_0)\} > 0$. For $\texttt{BanLast}(K, m)$ in case of $d > (2K+1)m$ it holds that

$$N_0 = 2 \text{ and } \delta = p(s, s, 2) = \left(\frac{C_{d-2Km}^m}{C_{d-Km}^m}\right)^K \cdot \left(\frac{1}{C_{d-Km}^m}\right)^K,$$

because the smallest probability is to return to state s in two steps.

• For $\texttt{KAWASAKI}(K, b, \pi_\Delta, m)$ from any given state, there exists a path to any other state in just one iteration, because probabilities to choose any set of coordinates ν are non-zero. Thus, the corresponding markov chain is indecomposable.

We focus on the case where $K = 1$ and that generalize analysis to accommodate larger values of K. Let us look at probabilities to move from ν_i to ν_j and from ν_j to ν_i. We show that both these probabilities correspond to random choice of the same indexes with the same distribution vector p, defined in 6, i.e. the probabilities are equal. For this case let us define ν as operator

$$\Psi_i(\overline{1,d}) := \nu_i,$$

i.e. operator chooses indexes that are in ν_i from $\overline{1,d}$. And

$$\Phi(p, \Psi_i) := \mathbb{P}\{\text{choose } \nu_i \text{ with distribution vector } p\}.$$

According to 6, probability to move from ν_i to ν_j equals a probability to choose indexes ν_j with distribution

$$p_i = \pi_\Delta(\widetilde{p}_i),$$

where

$$\widetilde{p}_i^k = \begin{cases} 1/bd & \text{if } k \in \nu_i \\ 1/d & \text{if } k \notin \nu_i \end{cases},$$

i.e.

$$p_{ij} = \Phi(p_i, \Psi_j).$$

By the definition of Φ, for arbitrary permutation ϕ and index choice Ψ holds

$$\Phi(\phi(p), \Psi \circ \phi) = \Phi(p, \Psi).$$

Now we point out that for arbitrary ν_i and ν_j exists permutation ϕ_{ij}, such that

$$\Psi_j \circ \phi_{ij} = \Psi_i.$$

For such permutation holds $\phi_{ij}(\widetilde{p}_i) = \widetilde{p}_j$, i.e. the permutations moves indexes from ν_i to indexes from ν_j. Then we need to use the property of π_Δ to get the same equality for p_i, p_j:

$$\phi_{ij}(p_i) = \phi_{ij}(\pi_\Delta(\widetilde{p}_i)) = \pi_\Delta \phi_{ij}((\widetilde{p}_i)) = \pi_\Delta(p_j).$$

This allows us to write

$$p_{ij} = \Phi(p_i, \Psi_j) = \Phi(\phi_{ij}(p_i), \Psi_j \circ \phi_{ij}) = \Phi(p_j, \Psi_i) = p_{ji}.$$

Thus we get equality of probabilities to move from ν_j to ν_i and to opposite way.

Now we can easily generalize the proof for arbitrary K. All that is required is to consider, instead of the sets of indices ν, combinations of sets of indices that were chosen for transmission over the previous K steps. In this way, the number of states is increased, but the logic of reasoning remains unchanged.

• As was mentioned above, for KAWASAKI(K, b, π_Δ, m) $N_0 = 1$. We now compute $\delta := p(s, s, 1)$, where $s = \{\nu, ..., \nu\}$, where ν occurs K times. In this case probability to choose ν another K times is equal to $\mathbb{P}\{j \in \nu\}^{mK}$. And

$$\mathbb{P}\{j \in \nu\} = \min\left\{\pi_\Delta\left[\widetilde{p} := \Big(\underbrace{\frac{1/d}{b^K}, ..., \frac{1/d}{b^K}}_{m}, \underbrace{\frac{1/d}{1}, ..., \frac{1/d}{1}}_{d-m}\Big)^T\right]\right\}.$$

If we consider $(\pi_\Delta(\widetilde{p}))_j = |\widetilde{p}_j|/\|\widetilde{p}\|_1$, then, since $\|\widetilde{p}\|_1 = \frac{1}{db^k}(db^K - m(b^K - 1))$, it hold that $\delta = (db^K - m(b^K - 1))^{-mK}$. This finishes the proof. □

D MAIN LEMMAS

Lemma 8. *For any $i \in \overline{1, n}$, $\varepsilon > 0$, $\tau > \tau_{mix}(\varepsilon)$, $t > \tau$, for any $a^{t-\tau}, b^{t-\tau} \in \mathbb{R}^d$, such that if we fix all randomness up to step $t - \tau$, $a^{t-\tau}$ and $b^{t-\tau}$ become non-random, it holds that*

$$\mathbb{E}\left[\left\langle Q_t^i\left(a^{t-\tau}\right) - a^{t-\tau}, b^{t-\tau}\right\rangle\right] \leq \frac{\varepsilon d}{m}\mathbb{E}\left[\left\|a^{t-\tau}\right\| \cdot \left\|b^{t-\tau}\right\|\right].$$

Proof. We begin by using tower property:

$$\mathbb{E}\left[\left\langle Q_t^i\left(a^{t-\tau}\right) - a^{t-\tau}, b^{t-\tau}\right\rangle\right] = \mathbb{E}\left[\left\langle \mathbb{E}_{t-\tau}\left[Q_t^i\left(a^{t-\tau}\right) - a^{t-\tau}\right], b^{t-\tau}\right\rangle\right], \tag{5}$$

where $\mathbb{E}_{t-\tau}[\cdot]$ is the conditional expectation with fixed randomness of all steps up to $t - \tau$. Since on a step t we compress vector $a^{t-\tau}$ according to distribution π_t^i by the formula $Q_t^i(a^{t-\tau}) = d/m a^{t-\tau} \odot \mathbb{1}(\nu_t^i)$, where ν_t^i is some set of m coordinates : $\nu_t^i \subset \overline{1, d}$ and $\mathbb{1}(\nu_t^i)$ is vector with 1 on coordinates ν_t^i on 0 otherwise. Using this we can obtain:

$$\mathbb{E}_{t-\tau}\left[Q_t^i\left(a^{t-\tau}\right) - a^{t-\tau}\right] = \sum_{\widetilde{\nu}_i \in M}\left(\mathbb{P}_{t-\tau}\left\{\nu_t^i = \widetilde{\nu}_i\right\} - \frac{1}{C_d^m}\right) a^{t-\tau} \odot \mathbb{1}(\widetilde{\nu}_i)\frac{d}{m},$$

where M is set of all subsets of $\overline{1, d}$ of size m. This equality follows from the fact that $\sum_{\widetilde{\nu}_i \in M} a^{t-\tau} \odot \mathbb{1}(\widetilde{\nu}_i) = C_{d-1}^{m-1} a^{t-\tau}$ and $C_{d-1}^{m-1}/C_d^m = m/d$. Now with the help of Cauchy–Schwarz inequality A.1 we can estimate (5):

$$(5) \leq \mathbb{E}\left[\sum_{\widetilde{\nu}_i \in M} \left|\mathbb{P}_{t-\tau}\left\{\nu_t^i = \widetilde{\nu}_i\right\} - \frac{1}{C_d^m}\right| \left\|a^{t-\tau} \odot \mathbb{1}(\widetilde{\nu}_i)\right\| \frac{d}{m} \left\|b^{t-\tau}\right\|\right]. \tag{6}$$

Since $t > \tau$ and $\tau > \tau_{\text{mix}}(\varepsilon)$ it holds that $\left|\mathbb{P}_{t-\tau}\left\{\nu_t^i = \widetilde{\nu}_i\right\} - 1/C_d^m\right| \leq \varepsilon \cdot 1/C_d^m$, because stationary distribution of our Markov chain is uniform. Using the fact that $\|a^{t-\tau} \odot \mathbb{1}(\widetilde{\nu}_i)\| \leq \|a^{t-\tau}\|$ we can obtain:

$$(6) \leq \mathbb{E}\left[\sum_{\widetilde{\nu}_i \in M} \varepsilon \frac{1}{C_d^m} \left\|a^{t-\tau}\right\| \frac{d}{m} \left\|b^{t-\tau}\right\|\right] = \frac{\varepsilon d}{m} \mathbb{E}\left[\left\|a^{t-\tau}\right\| \cdot \left\|b^{t-\tau}\right\|\right].$$

This finishes the proof.

□

Lemma 9. *For any $i \in \overline{1,n}$, $\varepsilon > 0$, $\tau > \tau_{mix}(\varepsilon)$, $t > \tau$, for any $a^{t-\tau} \in \mathbb{R}^d$, such that if we fix all randomness up to step $t-\tau$, $a^{t-\tau}$ becomes non-random, it holds that*

$$\mathbb{E}\left[\left\|\mathbb{E}_{t-\tau}\left[Q_t^i(a^{t-\tau})\right] - a^{t-\tau}\right\|^2\right] \leq \frac{\varepsilon^2 d^2}{m^2} \mathbb{E}\left[\left\|a^{t-\tau}\right\|^2\right].$$

Proof. Using same notation as in the proof of Lemma 11 we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbb{E}_{t-\tau}\left[Q_t^i(a^{t-\tau})\right] - a^{t-\tau}\right\|^2\right] &= \mathbb{E}\left[\left\|\sum_{\widetilde{\nu}_i \in M}\left(\mathbb{P}_{t-\tau}\left\{\nu_t^i = \widetilde{\nu}_i\right\} - \frac{1}{C_d^m}\right) \frac{d}{m} a^{t-\tau} \odot \mathbb{1}(\widetilde{\nu}_i)\right\|^2\right] \\ &\leq \mathbb{E}\left[\frac{d^2}{m^2} C_d^m \sum_{\widetilde{\nu}_i \in M}\left(\left|\mathbb{P}_{t-\tau}\left\{\nu_t^i = \widetilde{\nu}_i\right\} - \frac{1}{C_d^m}\right|^2 \left\|a^{t-\tau} \odot \mathbb{1}(\widetilde{\nu}_i)\right\|^2\right)\right].\end{aligned}$$

Since $t > \tau$ and $\tau > \tau_{\text{mix}}(\varepsilon)$ it holds that $\left|\mathbb{P}_{t-\tau}\left\{\nu_t^i = \widetilde{\nu}_i\right\} - 1/C_d^m\right| \leq \varepsilon \cdot 1/C_d^m$, because stationary distribution of our Markov chain is uniform. Using the fact that $\|a^{t-\tau} \odot \mathbb{1}(\widetilde{\nu}_i)\| \leq \|a^{t-\tau}\|$ we can obtain:

$$\mathbb{E}\left[\left\|\mathbb{E}_{t-\tau}\left[Q_t^i(a^{t-\tau})\right] - a^{t-\tau}\right\|^2\right] \leq \frac{\varepsilon^2 d^2}{m^2} \mathbb{E}\left[\left\|a^{t-\tau}\right\|^2\right].$$

This finishes the proof.

□

Lemma 10. *For any $i \in \overline{1,n}$ and $a \in \mathbb{R}^d$ it holds that*

$$\left\|Q^i(a)\right\|^2 \leq \frac{d^2}{m^2} \|a\|^2 \quad \textit{and} \quad \left\|Q^i(a) - a\right\|^2 \leq 4\frac{d^2}{m^2} \|a\|^2.$$

Proof. Consider the first inequality. Since $Q^i(a) = d/m\, a \odot \mathbb{1}(\nu^i)$, then $\left\|Q^i(a)\right\| \leq d/m \|a\|$, therefore

$$\left\|Q^i(a)\right\|^2 \leq \frac{d^2}{m^2} \|a\|^2.$$

Consider the second inequality. Using Fenchel-Young inequality A.2 with $\beta = 1$ we can estimate

$$\left\|Q^i(a) - a\right\|^2 \leq 2\left\|Q^i(a)\right\|^2 + 2\|a\|^2 \leq 2\left(\frac{d^2}{m^2} + 1\right)\|a\|^2 \leq 4\frac{d^2}{m^2}\|a\|^2.$$

This finishes the proof. □

Corollary 11. *For any $i \in \overline{1,n}$, $\varepsilon > 0$, $\tau > \tau_{mix}(\varepsilon)$, $t > \tau$, for any $a^t, b^t \in \mathbb{R}^d$, such that if we fix all randomness up to step t, a^t and b^t become non-random. And for any $\hat{a}^{t-\tau}, \hat{b}^{t-\tau}$, such that if we fix all randomness up to step $t-\tau$, $\hat{a}^{t-\tau}$ and $\hat{b}^{t-\tau}$ become non-random, it holds that*

$$2\left|\mathbb{E}\left[\left\langle Q_t^i\left(a^t\right) - a^t, b^t\right\rangle\right]\right| \leq \frac{\varepsilon d}{m\beta_0}\mathbb{E}\left[\left\|\hat{a}^{t-\tau}\right\|^2\right] + \frac{\varepsilon d\beta_0}{m}\mathbb{E}\left[\left\|\hat{b}^{t-\tau}\right\|^2\right] + \frac{1}{\beta_2}\mathbb{E}\left[\left\|b^t\right\|^2\right],$$
$$+\left(\frac{1}{\beta_1}+\frac{1}{\beta_3}\right)\mathbb{E}\left[\left\|b^t-\hat{b}^{t-\tau}\right\|^2\right] + 4\frac{d^2}{m^2}\beta_3\mathbb{E}\left[\left\|a^t\right\|^2\right] + 4\frac{d^2\left(\beta_1+\beta_2\right)}{m^2}\mathbb{E}\left[\left\|a^t-\hat{a}^{t-\tau}\right\|^2\right]$$

where $\beta_0, \beta_1, \beta_2, \beta_3 > 0$.

Proof. Using straightforward algebra we obtain

$$\begin{aligned}\mathbb{E}\left[\left\langle Q_t^i\left(a^t\right) - a^t, b^t\right\rangle\right] &= \mathbb{E}\left[\left\langle Q_t^i\left(\hat{a}^{t-\tau}\right) - \hat{a}^{t-\tau}, \hat{b}^{t-\tau}\right\rangle\right] \\ &- \mathbb{E}\left[\left\langle Q_t^i\left(a^t-\hat{a}^{t-\tau}\right) - a^t + \hat{a}^{t-\tau}, b^t - \hat{b}^{t-\tau}\right\rangle\right] \\ &+ \mathbb{E}\left[\left\langle Q_t^i\left(a^t-\hat{a}^{t-\tau}\right) - a^t + \hat{a}^{t-\tau}, b^t\right\rangle\right] \\ &+ \mathbb{E}\left[\left\langle Q_t^i\left(a^t\right) - a^t, b^t - \hat{b}^{t-\tau}\right\rangle\right].\end{aligned}$$

Using Lemma 8 with $a^{t-\tau} = \hat{a}^{t-\tau}, b^{t-\tau} = \hat{b}^{t-\tau}$ and Fenchel-Young inequality A.2 with $\beta_1, \beta_2, \beta_3 > 0$ we obtain:

$$\begin{aligned}2\left|\mathbb{E}\left[\left\langle Q_t^i\left(a^t\right) - a^t, b^t\right\rangle\right]\right| &\leq 2\frac{\varepsilon d}{m}\mathbb{E}\left[\left\|\hat{a}^{t-\tau}\right\| \cdot \left\|\hat{b}^{t-\tau}\right\|\right] \\ &+ \beta_1\mathbb{E}\left[\left\|Q_t^i\left(a^t-\hat{a}^{t-\tau}\right) - a^t + \hat{a}^{t-\tau}\right\|^2\right] + \frac{1}{\beta_1}\mathbb{E}\left[\left\|b^t-\hat{b}^{t-\tau}\right\|^2\right] \\ &+ \beta_2\mathbb{E}\left[\left\|Q_t^i\left(a^t-\hat{a}^{t-\tau}\right) - a^t + \hat{a}^{t-\tau}\right\|^2\right] + \frac{1}{\beta_2}\mathbb{E}\left[\left\|b^t\right\|^2\right] \\ &+ \beta_3\mathbb{E}\left[\left\|Q_t^i\left(a^t\right) - a^t\right\|^2\right] + \frac{1}{\beta_3}\mathbb{E}\left[\left\|b^t-\hat{b}^{t-\tau}\right\|^2\right].\end{aligned}$$

Using Lemma 10 and Fenchel-Young inequality A.2 with $\beta_0 > 0$ we obtain

$$\begin{aligned}2\left|\mathbb{E}\left[\left\langle Q_t^i\left(a^t\right) - a^t, b^t\right\rangle\right]\right| &\leq \frac{\varepsilon d}{m\beta_0}\mathbb{E}\left[\left\|\hat{a}^{t-\tau}\right\|^2\right] + \frac{\varepsilon d\beta_0}{m}\mathbb{E}\left[\left\|\hat{b}^{t-\tau}\right\|^2\right] \\ &+ 4\frac{d^2}{m^2}\left(\beta_1+\beta_2\right)\mathbb{E}\left[\left\|a^t-\hat{a}^{t-\tau}\right\|^2\right] + \left(\frac{1}{\beta_1}+\frac{1}{\beta_3}\right)\mathbb{E}\left[\left\|b^t-\hat{b}^{t-\tau}\right\|^2\right] \\ &+ 4\frac{d^2}{m^2}\beta_3\mathbb{E}\left[\left\|a^t\right\|^2\right] + \frac{1}{\beta_2}\mathbb{E}\left[\left\|b^t\right\|^2\right].\end{aligned}$$

This finishes the proof. □

Lemma 12. *Assume 4, then for any $x \in \mathbb{R}^d$ it holds that*

$$\frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(x)\right\|^2 \leq 2(\delta^2+1)\left\|\nabla f(x)\right\|^2 + 2\sigma^2.$$

Proof. Using straightforward algebra and Fenchel-Young inequality A.2 with $\beta = 1$ we obtain

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x)\|^2 &\leq \frac{2}{n}\sum_{i=1}^{n}\|\nabla f_i(x) - \nabla f(x)\|^2 + 2\|\nabla f(x)\|^2 \\ &\leq 2(\delta^2+1)\|\nabla f(x)\|^2 + 2\sigma^2.\end{aligned}$$

The last inequity follows from 4. This finishes the proof. □

E EXTENSIONS FOR THEOREM 3

E.1 FULL VERSION OF THEOREM 3

Theorem 13 (Convergence of MQSGD (Algorithm 1), extension of 3)**.** *Consider Assumptions 1, 4 and 5. Let problem* (1) *be solved by Algorithm 1.*

- *For any $\varepsilon > 0$, $\gamma > 0$, $\tau > \tau_{mix}(\varepsilon)$ and $T > \tau$ satisfying*

$$\gamma \lesssim \frac{m^2}{d^2 L(\delta^2+1)\tau} \quad \text{and} \quad \varepsilon \lesssim \frac{m^2}{d^2(\delta^2+1)},$$

it holds that

$$\mathbb{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right] = \mathcal{O}\left(\frac{F_\tau}{\gamma T} + \frac{\gamma L \tau d^2}{m^2}\sigma^2\right),$$

where $\widehat{x}^T$ is chosen uniformly from $\{x^t\}_{t=0}^T$.

- *If f additionally verifies the PL-condition (Assumption 3), then for any $\varepsilon > 0$, $\gamma > 0$, $\tau > \tau_{mix}(\varepsilon)$ and $T > \tau$ satisfying*

$$\gamma \lesssim \frac{m^2}{Ld^2\tau(\delta^2+1)} \quad \text{and} \quad \varepsilon = \sqrt{\gamma L \tau} \lesssim \frac{m}{d\sqrt{\delta^2+1}},$$

it holds that

$$F_T = \mathcal{O}\left(\left(1 - \frac{\mu\gamma}{12}\right)^{T-\tau} F_\tau + \frac{\gamma d^2 L \tau}{\mu m^2}\sigma^2\right).$$

Here we use a notation $F_t := \mathbb{E}\left[f(x^t) - f(x^*)\right]$.

E.2 FULL VERSION OF COROLLARY 4

Corollary 14 (Step tuning for Theorem 3, extension of Corollary 4)**.**

- *Under the conditions of Theorem 3 in the non-convex case, choosing γ as*

$$\gamma \lesssim \frac{m}{d\sqrt{L\tau}} \min\left\{\frac{m}{d(\delta^2+1)\sqrt{L\tau}}\,;\, \sqrt{\frac{F_\tau}{T\sigma^2}},\right\},$$

in order to achieve ϵ-approximate solution (in terms of $\mathbb{E}\left[\left\|\nabla f(x^T)\right\|^2\right] \leq \epsilon^2$) it takes

$$\mathcal{O}\left(\frac{L\tau d^2}{m^2} F_\tau \left(\frac{\delta^2+1}{\epsilon^2} + \frac{\sigma^2}{\epsilon^4}\right)\right) \text{ iterations of Algorithm 1.}$$

- *Under the conditions of Theorem 3 in the PL-condition (Assumption 3) case, choosing γ as*

$$\gamma \lesssim \min\left\{\frac{m^2}{Ld^2\tau(\delta^2+1)}\,;\,\frac{\log\left(\max\left\{2;\frac{\mu^2m^2F_\tau T}{d^2L\tau\sigma^2}\right\}\right)}{\mu T}\right\},$$

in order to achieve ϵ-approximate solution (in terms of $\mathbb{E}\left[f(x^t)-f(x^*)\right]\leq\epsilon$) it takes

$$\mathcal{O}\left(\frac{d^2L\tau}{m^2\mu}\left((\delta^2+1)\log\left(\frac{1}{\epsilon}\right)+\frac{\sigma^2}{\mu\epsilon}\right)\right) \text{ iterations of Algorithm 1.}$$

E.3 PROOF OF THEOREM 3, NON-CONVEX CASE

Proof. Denoting $F_t := \mathbb{E}\left[f(x^t)-f(x^*)\right]$, we have using L-smoothness:

$$F_{t+1}-F_t \leq -\gamma\mathbb{E}\left[\left\langle\frac{1}{n}\sum_{i=1}^n Q_t^i(\nabla f_i(x^t)),\nabla f(x^t)\right\rangle\right]+\frac{\gamma^2L}{2}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n Q_t^i(\nabla f_i(x^t))\right\|^2\right]. \tag{7}$$

Consider $-\gamma\mathbb{E}\left[\left\langle\frac{1}{n}\sum_{i=1}^n Q_t^i(\nabla f_i(x^t)),\nabla f(x^t)\right\rangle\right]$. Using straightforward algebra: $\pm\nabla f_i(x^{t-\tau})$ and $\pm\nabla f(x^{t-\tau})$ we can re-write this term:

$$\begin{aligned}
&-\gamma\,\mathbb{E}\left[\left\langle\frac{1}{n}\sum_{i=1}^n Q_t^i(\nabla f_i(x^t)),\nabla f(x^t)\right\rangle\right]\\
&=\underbrace{-\gamma\mathbb{E}\left[\left\langle\frac{1}{n}\sum_{i=1}^n Q_t^i(\nabla f_i(x^{t-\tau})),\nabla f(x^{t-\tau})\right\rangle\right]}_{①}\\
&\quad\underbrace{-\gamma\mathbb{E}\left[\left\langle\frac{1}{n}\sum_{i=1}^n Q_t^i(\nabla f_i(x^t)),\nabla f(x^t)-\nabla f(x^{t-\tau})\right\rangle\right]}_{②}\\
&\quad\underbrace{-\gamma\mathbb{E}\left[\left\langle\frac{1}{n}\sum_{i=1}^n Q_t^i(\nabla f_i(x^t)-\nabla f_i(x^{t-\tau})),\nabla f(x^{t-\tau})\right\rangle\right]}_{③}.
\end{aligned}$$

Consider ①. Using straightforward algebra, tower property, Lemmas 9 and 12 we obtain

$$
\begin{aligned}
① &= -\gamma\mathbb{E}\left[\left\langle \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{t-\tau}\left[Q_t^i(\nabla f_i(x^{t-\tau}))\right], \nabla f(x^{t-\tau})\right\rangle\right] \\
&= -\frac{\gamma}{2}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{t-\tau}\left[Q_t^i(\nabla f_i(x^{t-\tau}))\right]\right\|^2\right] \\
&\quad + \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^{t-\tau}) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{t-\tau}\left[Q_t^i(\nabla f_i(x^{t-\tau}))\right]\right\|^2\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^{t-\tau})\right\|^2\right] \\
&\leq \frac{\gamma}{2}\varepsilon^2\frac{d^2}{m^2}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_i(x^{t-\tau})\right\|^2\right] - \frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x^{t-\tau})\right\|^2\right] \\
&\leq \gamma\left(\varepsilon^2\frac{d^2}{m^2}(\delta^2+1) - \frac{1}{2}\right)\mathbb{E}\left[\left\|\nabla f(x^{t-\tau})\right\|^2\right] + \gamma\varepsilon^2\frac{d^2}{m^2}\sigma^2 \\
&\leq -\frac{\gamma}{4}\mathbb{E}\left[\left\|\nabla f(x^{t-\tau})\right\|^2\right] + \gamma\varepsilon^2\frac{d^2}{m^2}\sigma^2.
\end{aligned} \tag{8}
$$

The last inequality follows from the fact, that

$$
\varepsilon \leq \frac{m}{2d\sqrt{\delta^2+1}}.
$$

Consider ②. Using Cauchy-Schwarz A.1 and Fenchel-Young A.2 with $\beta = 1$ inequalities we obtain

$$
\begin{aligned}
② &\leq \mathbb{E}\left[\left\|-\frac{\gamma}{n}\sum_{i=1}^{n}Q_t^i(\nabla f_i(x^t))\right\|\left\|\nabla f(x^t) - \nabla f(x^{t-\tau})\right\|\right] \\
&\leq \gamma L\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}Q_t^i(\nabla f_i(x^t))\right\|\left\|x^t - x^{t-\tau}\right\|\right] \\
&= \gamma^2 L\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}Q_t^i(\nabla f_i(x^t))\right\|\left\|\sum_{s=t-\tau}^{t-1}\frac{1}{n}\sum_{i=1}^{n}Q_s^i(\nabla f_i(x^s))\right\|\right] \\
&\leq \frac{\gamma^2 L}{2}\left(\tau\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}Q_t^i(\nabla f_i(x^t))\right\|^2\right] + \sum_{s=t-\tau}^{t-1}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}Q_s^i(\nabla f_i(x^s))\right\|^2\right]\right).
\end{aligned} \tag{9}
$$

Third equality holds since $x^t - x^{t-\tau} = \gamma\sum_{s=t-\tau}^{t-1}\frac{1}{n}\sum\limits_{i=1}^{n}Q_s^i(\nabla f_i(x^s))$. Consider ③. Using Cauchy-Schwarz A.1 and Fenchel-Young A.2 with $\beta = m/d$ inequalities we obtain

$$
\begin{aligned}
③ &\leq \mathbb{E}\left[\left\|-\frac{\gamma}{n}\sum_{i=1}^{n}Q_t^i(\nabla f_i(x^t) - \nabla f_i(x^{t-\tau}))\right\|\left\|\nabla f(x^{t-\tau})\right\|\right] \\
&\leq \gamma L\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}Q_t^i(\nabla f_i(x^t) - \nabla f_i(x^{t-\tau})\right\|\left\|\nabla f(x^{t-\tau})\right\|\right] \\
&\leq \gamma^2 L\frac{d}{m}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}Q_t^i(\nabla f_i(x^t))\right\|\left\|\sum_{s=t-\tau}^{t-1}\frac{1}{n}\sum_{i=1}^{n}Q_s^i(\nabla f_i(x^s))\right\|\right] \\
&\leq \frac{\gamma^2 L}{2}\left(\sum_{s=t-\tau}^{t-1}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}Q_s^i(\nabla f_i(x^s))\right\|^2\right] + \frac{d^2\tau}{m^2}\mathbb{E}\left[\left\|\nabla f(x^{t-\tau})\right\|^2\right]\right).
\end{aligned} \tag{10}
$$

Wrapping (7) - (10) up we obtain

$$
\begin{aligned}
F_{t+1} - F_t &\leq \frac{\gamma^2 L}{2}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} Q_t^i(\nabla f_i(x^t))\right\|^2\right] - \frac{\gamma}{4}\mathbb{E}\left[\left\|\nabla f(x^{t-\tau})\right\|^2\right] + \gamma\varepsilon^2\frac{d^2}{m^2}\sigma^2 \\
&\quad + \frac{\gamma^2 L}{2}\left(\tau\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} Q_t^i(\nabla f_i(x^t))\right\|^2\right] + \sum_{s=t-\tau}^{t-1}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} Q_s^i(\nabla f_i(x^s))\right\|^2\right]\right) \\
&\quad + \frac{\gamma^2 L}{2}\left(\sum_{s=t-\tau}^{t-1}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} Q_s^i(\nabla f_i(x^s))\right\|^2\right] + \frac{d^2\tau}{m^2}\mathbb{E}\left[\left\|\nabla f(x^{t-\tau})\right\|^2\right]\right) \\
&\leq \gamma^2 L\tau\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} Q_t^i(\nabla f_i(x^t))\right\|^2\right] + \gamma\varepsilon^2\frac{d^2}{m^2}\sigma^2 \\
&\quad + \gamma^2 L\sum_{s=t-\tau}^{t-1}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} Q_s^i(\nabla f_i(x^s))\right\|^2\right] + \left(\frac{\gamma^2 L\tau d^2}{2m^2} - \frac{\gamma}{4}\right)\mathbb{E}\left[\left\|\nabla f(x^{t-\tau})\right\|^2\right].
\end{aligned}
$$

Using Lemma 12 we obtain

$$
\begin{aligned}
F_{t+1} - F_t &\leq \frac{2d^2\gamma^2 L\tau}{m^2}\left((\delta^2+1)\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + \sigma^2\right) + \left(\frac{\gamma^2 L\tau d^2}{2m^2} - \frac{\gamma}{4}\right)\mathbb{E}\left[\left\|\nabla f(x^{t-\tau})\right\|^2\right] \\
&\quad + \frac{2d^2\gamma^2 L}{m^2}\sum_{s=t-\tau}^{t-1}\left((\delta^2+1)\mathbb{E}\left[\left\|\nabla f(x^s)\right\|^2\right] + \sigma^2\right) + \gamma\varepsilon^2\frac{d^2}{m^2}\sigma^2 \\
&= \frac{2d^2\gamma^2 L(\delta^2+1)\tau}{m^2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + \frac{2d^2\gamma^2 L(\delta^2+1)}{m^2}\sum_{s=t-\tau}^{t-1}\mathbb{E}\left[\left\|\nabla f(x^s)\right\|^2\right] \\
&\quad + \left(\frac{\gamma^2 L\tau d^2}{2m^2} - \frac{\gamma}{4}\right)\mathbb{E}\left[\left\|\nabla f(x^{t-\tau})\right\|^2\right] + \frac{\gamma d^2}{m^2}\left(4\gamma L\tau + \varepsilon^2\right)\sigma^2.
\end{aligned} \tag{11}
$$

Summing (11) from $t = \tau$ to $t = T$ and using the fact that $\varepsilon^2 \leq \gamma L\tau$ and $1 + \delta^2 \geq 1$ we obtain

$$
\begin{aligned}
\sum_{t=\tau}^{T}\frac{\gamma}{4}\mathbb{E}\left[\left\|\nabla f(x^{t-\tau})\right\|^2\right] &\leq F_\tau + \frac{2d^2\gamma^2 L(\delta^2+1)}{m^2}\left(\tau\sum_{t=\tau}^{T}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right]\right. \\
&\quad \left. + \sum_{t=\tau}^{T}\sum_{s=t-\tau}^{t-1}\mathbb{E}\left[\left\|\nabla f(x^s)\right\|^2\right] + \tau\sum_{t=\tau}^{T}\mathbb{E}\left[\left\|\nabla f(x^{t-\tau})\right\|^2\right]\right) + \sum_{t=\tau}^{T} 5\frac{\gamma^2 L\tau d^2}{m^2}\sigma^2.
\end{aligned}
$$

Since $\sum_{t=\tau}^{T}\sum_{s=t-\tau}^{t-1}\mathbb{E}\left[\left\|\nabla f(x^s)\right\|^2\right] \leq \tau\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right]$, we get

$$
\gamma\sum_{t=0}^{T-\tau}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \leq 4F_\tau + \frac{24d^2\gamma^2 L(\delta^2+1)\tau}{m^2}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + 20\sum_{t=\tau}^{T}\frac{\gamma^2 L\tau d^2}{m^2}\sigma^2.
$$

Taking

$$
\gamma \leq \frac{m^2}{48d^2 L(\delta^2+1)\tau},
$$

we obtain

$$\gamma \sum_{t=0}^{T-\tau} \mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \leq 8F_\tau + \frac{48d^2\gamma^2 L(\delta^2+1)\tau}{m^2} \sum_{t=T-\tau}^{T} \mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + 40\sum_{t=\tau}^{T} \frac{\gamma^2 L\tau d^2}{m^2}\sigma^2. \tag{12}$$

We now prove that for any $t \geq 0$, we have

$$\sup_{t\leq s\leq t+\tau} \left\{\mathbb{E}\left[\|\nabla f(x^s)\|^2\right]\right\} \leq 4\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + 8L^2\gamma^2\tau^2\frac{d^2}{m^2}\sigma^2.$$

For $t \leq s \leq t+\tau$ it holds that

$$\begin{aligned}
\mathbb{E}\left[\|\nabla f(x^s)\|^2\right] &\leq 2\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + \mathbb{E}\left[\left\|\nabla f(x^s) - \nabla f(x^t)\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + 2L^2\gamma^2\mathbb{E}\left[\left\|\sum_{r=t}^{s-1}\frac{1}{n}\sum_{i=1}^{n} Q_r^i(\nabla f_i(x^r))\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + 2L^2\gamma^2\tau\frac{d^2}{m^2}\sum_{r=t}^{s-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\nabla f_i(x^r)\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + 4L^2\gamma^2\tau\frac{d^2}{m^2}\sum_{r=t}^{s-1}\left((\delta^2+1)\mathbb{E}\left[\|\nabla f(x^r)\|^2\right] + \sigma^2\right) \\
&\leq 2\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + 4L^2\gamma^2\tau^2\frac{d^2}{m^2}\left((\delta^2+1)\sup_{t\leq s\leq t+\tau}\left\{\mathbb{E}\left[\|\nabla f(x^s)\|^2\right]\right\} + \sigma^2\right).
\end{aligned}$$

Since

$$\gamma \leq \frac{m}{\sqrt{8}dL\sqrt{\delta^2+1}\tau},$$

it holds that

$$\sup_{t\leq s\leq t+\tau} \left\{\mathbb{E}\left[\|\nabla f(x^s)\|^2\right]\right\} \leq 4\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + 8L^2\gamma^2\tau^2\frac{d^2}{m^2}\sigma^2.$$

Using this (12) takes form

$$\begin{aligned}
\gamma \sum_{t=0}^{T-\tau} \mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \leq 8F_\tau + \frac{192d^2\gamma^2 L(\delta^2+1)\tau}{m^2} \sum_{t=T-2\tau}^{T-\tau} \mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \\
+ 384L^3\gamma^4\tau^3\frac{d^4}{m^4}(\delta^2+1)\sigma^2 + 40\sum_{t=\tau}^{T}\frac{\gamma^2 L\tau d^2}{m^2}\sigma^2.
\end{aligned}$$

Taking

$$\gamma \leq \frac{m}{384dL\sqrt{\delta^2+1}\tau},$$

and dividing both sides of the inequality by $T-\tau$, we obtain

$$\frac{1}{T-\tau}\sum_{t=0}^{T-\tau}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \leq 16\frac{F_\tau}{\gamma(T-\tau)} + 80\frac{\gamma^2 L\tau d^2}{m^2}\sigma^2.$$

Therefore, if $\widehat{x}^T$ is chosen uniformly from $\{x^t\}_{t=0}^{T-1}$, then it holds that

$$\mathbb{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right] \leq 16\frac{F_\tau}{\gamma T} + 80\frac{\gamma^2 L\tau d^2}{m^2}\sigma^2.$$

This finishes the proof. □

E.4 PROOF OF THEOREM 3, UNDER PL-CONDITION

Proof. We start from (11):

$$F_{t+1} - F_t = \frac{2d^2\gamma^2 L(\delta^2+1)\tau}{m^2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + \frac{2d^2\gamma^2 L(\delta^2+1)}{m^2}\sum_{s=t-\tau}^{t-1}\mathbb{E}\left[\left\|\nabla f(x^s)\right\|^2\right]$$
$$+ \left(\frac{\gamma^2 L\tau d^2}{2m^2} - \frac{\gamma}{4}\right)\mathbb{E}\left[\left\|\nabla f(x^{t-\tau})\right\|^2\right] + \frac{\gamma d^2}{m^2}\left(4\gamma L\tau + \varepsilon^2\right)\sigma^2.$$

If f satisfies PL-inequality (Assumption 3), then $-\mathbb{E}\left[\left\|\nabla f(x^{t-\tau}\right\|^2\right] \leq -2\mu F_{t-\tau}$, so that, for some $0 < \alpha < 1$ we obtain

$$F_{t+1} - F_t = \frac{2d^2\gamma^2 L(\delta^2+1)\tau}{m^2}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] + \frac{2d^2\gamma^2 L(\delta^2+1)}{m^2}\sum_{s=t-\tau}^{t-1}\mathbb{E}\left[\left\|\nabla f(x^s)\right\|^2\right]$$
$$+ \left(\frac{\gamma^2 L\tau d^2}{2m^2} - \frac{(1-\alpha)\gamma}{4}\right)\mathbb{E}\left[\left\|\nabla f(x^{t-\tau})\right\|^2\right] \tag{13}$$
$$- \frac{\alpha\gamma\mu}{2}F_{t-\tau} + \frac{\gamma d^2}{m^2}\left(4\gamma L\tau + \varepsilon^2\right)\sigma^2.$$

For $t \geq 0$, let $p_t = p^t$ and $p = (1 - \alpha\mu\gamma/4)^{-1}$. We multiply the above expression by p_t and sum for $t < T$, hoping for cancellations. Using PL-condition (Assumption 3), for $T \geq \tau$ we obtain

$$\begin{aligned}
\sum_{t=\tau}^{T-1} p_{t+1}\left(F_t - F_{t+1} - \frac{\alpha\gamma\mu}{4}F_{t-\tau}\right) &= \sum_{t=\tau}^{T-1} p_{t+1}\left[\left(1 - \frac{\alpha\gamma\mu}{4}\right)F_t - F_{t+1} + \frac{\alpha\gamma\mu}{4}(F_t - F_{t-\tau})\right] \\
&= \sum_{t=\tau}^{T-1} p_t F_t - \sum_{t=\tau+1}^{T} p_t F_t + \frac{\alpha\gamma\mu}{4}\sum_{t=\tau}^{T-1} p_{t+1}(F_t - F_{t-\tau}) \\
&\leq p_\tau F_\tau - p_T F_T + \frac{\alpha\gamma\mu}{4}\sum_{t=\tau}^{T-1} p_{t+1}F_t \\
&\quad - \frac{\alpha\gamma\mu p_\tau}{4}\sum_{t=0}^{T-1-\tau} p_{t+1}F_t \\
&\leq p_\tau F_\tau - p_T F_T + \frac{\alpha\gamma\mu}{4}\sum_{t=T-\tau}^{T-1} p_{t+1}F_t \\
&\leq p_\tau F_\tau - p_T F_T + \frac{\alpha\gamma}{8}\sum_{t=T-\tau}^{T-1} p_{t+1}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right].
\end{aligned}$$

For any $t \geq 0$ we use a notation $b_t := \mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right]$. We now handle b_t terms from (13).

$$-\sum_{t=\tau}^{T-1}\frac{(1-\alpha)\gamma}{4}p_{t+1}b_{t-\tau}+\gamma^2L\frac{d^2}{m^2}\sum_{t=\tau}^{T-1}p_{t+1}\left(2\tau(\delta^2+1)b_t+2(\delta^2+1)\sum_{s=t-\tau}^{t-1}b_s+\frac{\tau}{2}b_{t-\tau}\right). \tag{14}$$

If $p_t=p^t$, $p=(1-\alpha\mu\gamma/2)^{-1}$ and $\gamma=\gamma_1/\tau$, then, using the fact that $(1-a/x)^{-x}\leq 2e^a\leq 2e$ if $x\geq 2$ and $0\leq a\leq 1$, we can get that $1\geq p_\tau=(1-\mu\gamma_1/(2\tau))^{-\tau}\leq 2e^{\mu\gamma_1/2}\leq 2e\leq 6$. Then

$$\sum_{t=\tau}^{T}p_{t+1}\sum_{s=t-\tau}^{t-1}b_s\leq p^\tau\sum_{t=\tau}^{T}\sum_{s=t-\tau}^{t-1}p_{s+1}b_s\leq 6\tau\sum_{t=0}^{T}p_{t+1}b_t.$$

Now we can estimate (14):

$$\begin{aligned}(14)&\leq -\sum_{t=0}^{T-\tau-1}\frac{(1-\alpha)\gamma}{4}p_{t+1}b_t+\gamma^2L\frac{d^2\tau}{m^2}\left(2(\delta^2+1)\sum_{t=\tau}^{T-1}b_t+12(\delta^2+1)\sum_{t=0}^{T-1}b_t+3\sum_{t=0}^{T-\tau}b_t\right)\\&\leq -\sum_{t=0}^{T-\tau-1}p_{t+1}\gamma b_t\left(\frac{1-\alpha}{4}-17\gamma L\frac{d^2\tau(\delta^2+1)}{m^2}\right)+14\gamma^2L\frac{d^2\tau(\delta^2+1)}{m^2}\sum_{t=T-\tau}^{T-1}p_{t+1}b_t.\end{aligned}$$

Taking

$$\gamma\leq\frac{m^2(1-\alpha)}{136Ld^2\tau(\delta^2+1)\beta},$$

where $\beta\geq 1$, we obtain

$$(14)\leq -\frac{(1-\alpha)\gamma}{8}\sum_{t=0}^{T-\tau-1}p_{t+1}b_t+\frac{(1-\alpha)\gamma}{4\beta}\sum_{t=T-\tau}^{T-1}p_{t+1}b_t.$$

Now we can estimate (13):

$$\begin{aligned}0\leq p_\tau F_\tau-p_TF_T+\left(\frac{\alpha\gamma}{8}+\frac{(1-\alpha)\gamma}{4\beta}\right)\sum_{t=T-\tau}^{T-1}p_{t+1}b_t-\frac{(1-\alpha)\gamma}{8}\sum_{t=0}^{T-\tau-1}p_{t+1}b_t\\+\sum_{t=\tau}^{T-1}p_{t+1}\frac{\gamma d^2}{m^2}\left(4\gamma L\tau+\varepsilon^2\right)\sigma^2.\end{aligned} \tag{15}$$

Using that we proved in E.3 we have $b_t\leq 4b_{t-\tau}+8L^2\gamma^2\tau^2\frac{d^2}{m^2}\sigma^2$. Then, we can obtain

$$\begin{aligned}\gamma\left(\frac{\alpha}{8}+\frac{1-\alpha}{4\beta}\right)\sum_{t=T-\tau}^{T-1}p_{t+1}b_t\leq 24\gamma\left(\frac{\alpha}{8}+\frac{1-\alpha}{4\beta}\right)\sum_{t=T-2\tau}^{T-\tau-1}p_{t+1}b_t\\+48L^2\gamma^3\tau^3\frac{d^2}{m^2}\left(\frac{\alpha}{8}+\frac{1-\alpha}{4\beta}\right)\sigma^2.\end{aligned}$$

Taking $\alpha=1/6$ and $\beta=4$, we obtain

$$\frac{\alpha}{8}+\frac{1-\alpha}{4\beta}=\frac{1-\alpha}{8},$$

and (15) takes form

$$0 \leq p_\tau F_\tau - p_T F_T + 48L^2\gamma^3\tau^3\frac{d^2}{m^2}\sigma^2 + \sum_{t=\tau}^{T-1} p_{t+1}\frac{\gamma d^2}{m^2}\left(4\gamma L\tau + \varepsilon^2\right)\sigma^2. \tag{16}$$

Using the fact that

$$\sum_{t=\tau}^{T}\left(1-\frac{\alpha\mu\gamma}{2}\right)^{T-t} = \sum_{t=0}^{T-\tau}\left(1-\frac{\alpha\mu\gamma}{2}\right)^{t} \leq \sum_{t=0}^{+\infty}\left(1-\frac{\alpha\mu\gamma}{2}\right)^{t} = \frac{2}{\alpha\mu\gamma},$$

and taking

$$\gamma \leq \frac{m^2}{625Ld^2\tau(\delta^2+1)} \quad \text{and} \quad \varepsilon = \sqrt{\gamma L\tau} \leq \frac{m}{25d\sqrt{\delta^2+1}},$$

by dividing (16) by p_τ, we obtain

$$\mathbb{E}\left[f(x^T) - f(x^*)\right] \leq \left(1-\frac{\mu\gamma}{12}\right)^{T-\tau}\mathbb{E}\left[f(x^\tau) - f(x^*)\right] + 636\frac{\gamma d^2 L\tau}{\mu m^2}\sigma^2.$$

This finishes the proof.

□

F CONVERGENCE OF ALGORITHM 1 WITHOUT DATA SIMILARITY

Theorem 15 (Convergence of GD Algorithm 1 without data similarity)**.** *Consider Assumptions 1 and 2. Let problem* (1) *be solved by Algorithm 1. Then for any $\varepsilon > 0$, $\gamma > 0$, $\tau > \tau_{mix}(\varepsilon)$ and $T > \tau$ satisfying*

$$\gamma \leq \frac{m^2\sqrt{\mu}}{24d^2L^{3/2}\tau} \quad \textit{and} \quad \varepsilon \leq \frac{m\sqrt{\mu}}{24d}\min\left\{\frac{1}{L^{3/2}}; \sqrt{\mu}\right\},$$

it holds that

$$\mathbb{E}\left[\left\|x^{T+1} - x^*\right\|^2\right] \leq \left(1-\frac{\mu\gamma}{2}\right)^{T-\tau}\mathbb{E}\left[\left\|x^\tau - x^*\right\|^2\right] + \left(1-\frac{\mu\gamma}{2}\right)^{T}\Delta_\tau + 26\frac{\gamma d^2\tau}{\mu m^2}\sigma_*^2,$$

where

$$\Delta_\tau = \mathcal{O}\left(\frac{\gamma^2 d^2}{m^2}\sqrt{\frac{\mu}{L}}\sum_{t=0}^{\tau}\left[\tau\mathbb{E}\left[\left\|x^t - x^*\right\|^2\right] + 4L\mathbb{E}\left[f(x^t) - f(x^*)\right]\right]\right).$$

Proof of Theorem 15. We start by writing out step of the Algorithm 1:

$$\begin{aligned}\mathbb{E}\left[\left\|x^{t+1} - x^*\right\|^2\right] = \mathbb{E}\left[\left\|x^t - x^*\right\|^2\right] - 2\gamma\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{d}\left\langle Q_t^i\left(\nabla f_i(x^t)\right) - \nabla f_i(x^t), x^t - x^*\right\rangle\right] \\ - 2\gamma\mathbb{E}\left[\left\langle\nabla f(x^t), x^t - x^*\right\rangle\right] + \gamma^2\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}Q_t^i\left(\nabla f_i(x^t)\right)\right\|^2\right].\end{aligned} \tag{17}$$

Consider $\mathbb{E}\left[\left\langle Q_t^i\left(\nabla f_i(x^t)\right) - \nabla f_i(x^t), x^t - x^*\right\rangle\right]$. Using Corollary 11 with $a^t = \nabla f_i(x^t)$, $b^t = x^t - x^*$, $\hat{a}^{t-\tau} = \nabla f_i(x^{t-\tau})$ and $\hat{b}^{t-tau} = x^{t-\tau} - x^*$ we obtain

$$
\begin{aligned}
&2\,\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}|\langle Q_t^i\left(\nabla f_i(x^t)\right) - \nabla f_i(x^t), x^t - x^*\rangle|] \leq \frac{\varepsilon d}{m\beta_0}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_i(x^{t-\tau})\right\|^2\right] \\
&+ \frac{\varepsilon d\beta_0}{m}\mathbb{E}\left[\left\|x^{t-\tau} - x^*\right\|^2\right] + 4\frac{d^2L^2}{m^2}(\beta_1+\beta_2)\mathbb{E}\left[\left\|x^t - x^\tau\right\|^2\right] + \left(\frac{1}{\beta_1}+\frac{1}{\beta_3}\right)\mathbb{E}\left[\left\|x^t - x^\tau\right\|^2\right] \\
&+ 4\frac{d^2}{m^2}\beta_3\frac{1}{n}\sum_{i=1}^{d}\mathbb{E}\left[\left\|\nabla f_i(x^t)\right\|^2\right] + \frac{1}{\beta_2}\mathbb{E}\left[\left\|x^t - x^*\right\|^2\right].
\end{aligned} \tag{18}
$$

Using the fact that f_i are L-smooth, we can obtain:

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(x^t)\right\|^2 &= \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(x^t) - \nabla f_i(x^*) + \nabla f_i(x^*)\right\|^2 \\
&\leq \frac{2}{n}\sum_{i=1}^{n}\left\|\nabla f_i(x^t) - \nabla f_i(x^*)\right\|^2 + \frac{2}{n}\sum_{i=1}^{n}\left\|\nabla f_i(x^*)\right\|^2 \\
&\leq \frac{4L}{n}\sum_{i=1}^{n}\left(f_i(x^t) - f_i(x^*) - \left\langle\nabla f_i(x^*), x^t - x^*\right\rangle\right) + 2\sigma_*^2 \\
&= 4L(f(x^t) - f(x^*)) + 2\sigma_*^2,
\end{aligned} \tag{19}
$$

where we use a notation $\sigma_*^2 := \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(x^*)\right\|^2$. Now we can estimate (18):

$$
\begin{aligned}
(18) \leq\ & \frac{2\varepsilon d}{m\beta_0}(2L\mathbb{E}\left[f(x^{t-\tau}) - f(x^*)\right] + \sigma_*^2) + \frac{\varepsilon d\beta_0}{m}\mathbb{E}\left[\left\|x^{t-\tau} - x^*\right\|^2\right] \\
&+ \left(4\frac{d^2L^2}{m^2}(\beta_1+\beta_2) + \frac{1}{\beta_1} + \frac{1}{\beta_3}\right)\mathbb{E}\left[\left\|-\gamma\sum_{s=t-\tau}^{t-1}\frac{1}{n}\sum_{i=1}^{n}Q_s^i\left(\nabla f_i(x^s)\right)\right\|^2\right] \\
&+ 8\frac{d^2}{m^2}\beta_3(2L\mathbb{E}\left[f(x^t) - f(x^*)\right] + \sigma_*^2) + \frac{1}{\beta_2}\mathbb{E}\left[\left\|x^t - x^*\right\|^2\right].
\end{aligned}
$$

Now we can estimate (17). Using Lemma 10 and Assumption 2 we can obtain

$$
\begin{aligned}
\mathbb{E}\left[\left\|x^{t+1} - x^*\right\|^2\right] \leq\ & \left(1 - \mu\gamma + \frac{\gamma}{\beta_2}\right)\mathbb{E}\left[\left\|x^t - x^*\right\|^2\right] + \frac{\varepsilon d\beta_0\gamma}{m}\mathbb{E}\left[\left\|x^{t-\tau} - x^*\right\|^2\right] \\
&+ 4L\mathbb{E}\Bigg[\frac{\varepsilon d\gamma}{m\beta_0}(f(x^{t-\tau}) - f(x^*)) + 4\frac{d^2\beta_3\gamma}{m^2}(f(x^t) - f(x^*)) \\
&+ \left(4\frac{d^2L^2}{m^2}(\beta_1+\beta_2) + \frac{1}{\beta_1} + \frac{1}{\beta_3}\right)\frac{\gamma^3\tau d^2}{m^2}\sum_{s=t-\tau}^{t-1}(f(x^s) - f(x^*)) \\
&+ \frac{\gamma^2 d^2}{m^2}(f(x^t) - f(x^*)) - \frac{\gamma}{2L}(f(x^t) - f(x^*))\Bigg] \\
&+ 2\left[\frac{\varepsilon d\gamma}{m\beta_0} + 4\frac{d^2\beta_3\gamma}{m^2} + \left(4\frac{d^2L^2}{m^2}(\beta_1+\beta_2) + \frac{1}{\beta_1} + \frac{1}{\beta_3}\right)\frac{\gamma^3\tau^2 d^2}{m^2} + \frac{\gamma^2 d^2}{m^2}\right]\sigma_*^2.
\end{aligned} \tag{20}
$$

Taking $\beta_0 = \beta_1 = 1, \beta_3 = \gamma, \beta_2 = 4/\mu$ and using fact, that $\varepsilon \leq \gamma\tau d/m$ inequality (20) takes form

$$\begin{aligned}\mathbb{E}\left[\left\|x^{t+1} - x^*\right\|^2\right] \leq \left(1 - \frac{3}{4\mu\gamma}\right)\mathbb{E}\left[\left\|x^t - x^*\right\|^2\right] + \frac{\varepsilon d\beta_0\gamma}{m}\mathbb{E}\left[\left\|x^{t-\tau} - x^*\right\|^2\right] \\ + 4L\mathbb{E}\left[\frac{\varepsilon d\gamma}{m\beta_0}(f(x^{t-\tau}) - f(x^*)) + 5\frac{d^2\gamma^2}{m^2}(f(x^t) - f(x^*))\right. \\ \left. + 20\frac{d^4L^2}{m^4}\frac{\gamma^3\tau}{\mu}\sum_{s=t-\tau}^{t-1}(f(x^s) - f(x^*)) - \frac{\gamma}{2L}(f(x^t) - f(x^*))\right] \\ + 4\frac{d^2\gamma^2\tau}{m^2}\left[3 + 10\frac{d^2L^2}{m^2}\frac{\gamma}{\mu}\right]\sigma_*^2.\end{aligned} \tag{21}$$

Let us perform the summation from $t = \tau$ to $t = T > \tau$ of equations (21) with coefficients p_k:

$$\sum_{t=\tau}^{T} p_t\mathbb{E}\left[\left\|x^{t+1} - x^*\right\|^2\right] \leq \sum_{t=\tau}^{T} p_t(1 - \frac{3\mu\gamma}{4})\mathbb{E}\left[\left\|x^t - x^*\right\|^2\right] \tag{22}$$

$$+ \sum_{t=\tau}^{T} p_t\frac{\gamma\varepsilon d}{m}\mathbb{E}\left[\left\|x^{t-\tau} - x^*\right\|^2\right] \tag{23}$$

$$+ \sum_{t=\tau}^{T} p_t 4L\left(\frac{\gamma\varepsilon d}{m} + 5\frac{\gamma^2d^2\tau}{m^2} - \frac{\gamma}{2L}\right)\mathbb{E}\left[f(x^t) - f(x^*)\right] \tag{24}$$

$$+ 20\sum_{t=\tau}^{T} p_t 4L\frac{d^4L^2}{m^4}\frac{\gamma^3\tau}{\mu}\sum_{s=t-\tau}^{t-1}\mathbb{E}\left[f(x^s) - f(x^*)\right] \tag{25}$$

$$+ \sum_{t=\tau}^{T} p_t 4\frac{d^2\gamma^2\tau}{m^2}\left[3 + 10\frac{d^2L^2}{m^2}\frac{\gamma}{\mu}\right]\sigma_*^2. \tag{26}$$

If $p_t = p^t$, $p = (1 - \mu\gamma/2)^{-1}$ and $\gamma = \gamma_1/\tau$, then, using the fact that $(1 - a/x)^{-x} \leq 2e^a \leq 2e$ if $x \geq 2$ and $0 \leq a \leq 1$, we can get that $p_\tau = (1 - \mu\gamma_1/(2\tau))^{-\tau} \leq 2e^{\mu\gamma_1/2} \leq 2e \leq 6$.

$$\sum_{t=\tau}^{T} p_t \sum_{s=t-\tau}^{t-1} a_s \leq p^\tau \sum_{t=\tau}^{T}\sum_{s=t-\tau}^{t-1} p_s a_s \leq 6\tau\sum_{t=0}^{T} p_t a_t.$$

Using this we can estimate (22):

$$\begin{aligned}&\sum_{t=\tau}^{T} p_t\mathbb{E}\left[\left\|x^{t+1} - x^*\right\|^2\right] \leq \sum_{t=\tau}^{T} p_t\left(1 - \mu\gamma + 6\frac{\gamma\varepsilon d}{m}\right)\mathbb{E}\left[\left\|x^t - x^*\right\|^2\right] \\ &+ \sum_{t=\tau}^{T} 4p_t L\left(\frac{\gamma\varepsilon d}{m} + 5\frac{\gamma^2d^2\tau}{m^2} + 120\frac{d^4L^2}{m^4}\frac{\gamma^3\tau^2}{\mu} - \frac{\gamma}{2L}\right)\mathbb{E}\left[f(x^t) - f(x^*)\right] \\ &+ 4\sum_{t=\tau}^{T} p_t\left[3 + 10\frac{d^2L^2}{m^2}\frac{\gamma}{\mu}\right]\sigma_*^2 + \sum_{t=0}^{\tau} p_{t+\tau}\frac{\gamma\varepsilon d}{m}\mathbb{E}\left[\left\|x^t - x^*\right\|^2\right] \\ &+ 80\sum_{t=0}^{\tau} p_{t+\tau}L\frac{d^4L^2}{m^4}\frac{\gamma^3\tau}{\mu}\mathbb{E}\left[f(x^t) - f(x^*)\right].\end{aligned} \tag{27}$$

Taking

$$\gamma \leq \frac{m^2\sqrt{\mu}}{24d^2L^{3/2}\tau} \quad \text{and} \quad \varepsilon = \min\left\{\frac{\gamma d\tau}{m}; \frac{\mu m}{24d}\right\} \leq \frac{m\sqrt{\mu}}{24d}\min\left\{\frac{1}{L^{3/2}}; \sqrt{\mu}\right\}.$$

We get

$$\frac{\gamma\varepsilon d}{m}+5\frac{\gamma^2d^2\tau}{m^2}+120\frac{d^4L^2}{m^4}\frac{\gamma^3\tau^2}{\mu}-\frac{\gamma}{2L}\leq 0 \quad \text{and} \quad 1-\frac{3\mu\gamma}{4}+6\frac{\gamma\varepsilon d}{m}=1-\frac{\mu\gamma}{2}.$$

Assume a notation

$$\begin{aligned}\Delta_\tau &:= \sum_{t=0}^{\tau} p_{t+\tau}\frac{\gamma\varepsilon d}{m}\mathbb{E}\left[\left\|x^t-x^*\right\|^2\right]+80\sum_{t=0}^{\tau}p_{t+\tau}L\frac{d^4L^2}{m^4}\frac{\gamma^3\tau}{\mu}\mathbb{E}\left[f(x^t)-f(x^*)\right]\\ &\leq 120\frac{\gamma^2d^2}{m^2}\sqrt{\frac{\mu}{L}}\sum_{t=0}^{\tau}\left(\tau\mathbb{E}\left[\left\|x^t-x^*\right\|^2\right]+4L\mathbb{E}\left[f(x^t)-f(x^*)\right]\right).\end{aligned}$$

Using the notation of Δ_τ, (27) takes form

$$\sum_{t=\tau}^{T}p_t\mathbb{E}\left[\left\|x^{t+1}-x^*\right\|^2\right]\leq\sum_{t=\tau}^{T}p_t\left(1-\frac{\mu\gamma}{2}\right)\mathbb{E}\left[\left\|x^t-x^*\right\|^2\right]+\sum_{t=\tau}^{T}13p_t\frac{\gamma^2d^2\tau}{m^2}\sigma_*^2+\Delta_\tau.$$

Using $p_t=p^t$ and $p=(1-\mu\gamma/2)^{-1}$ we can obtain:

$$\begin{aligned}\sum_{t=\tau}^{T}\left(1-\frac{\mu\gamma}{2}\right)^{-t}\mathbb{E}\left[\left\|x^{t+1}-x^*\right\|^2\right]\leq&\sum_{t=\tau}^{T}\left(1-\frac{\mu\gamma}{2}\right)^{-t+1}\mathbb{E}\left[\left\|x^t-x^*\right\|^2\right]\\&+\sum_{t=\tau}^{T}13\left(1-\frac{\mu\gamma}{2}\right)^{-t}\frac{\gamma^2d^2\tau}{m^2}\sigma_*^2+\Delta_\tau.\end{aligned}$$

The summed terms on the left and right sides are reduced, therefore this expression takes the form:

$$\begin{aligned}\left(1-\frac{\mu\gamma}{2}\right)^{-T}\mathbb{E}\left[\left\|x^{T+1}-x^*\right\|^2\right]\leq&\left(1-\frac{\mu\gamma}{2}\right)^{-\tau}\mathbb{E}\left[\left\|x^\tau-x^*\right\|^2\right]\\&+\sum_{t=\tau}^{T}13\left(1-\frac{\mu\gamma}{2}\right)^{-t}\frac{\gamma^2d^2\tau}{m^2}\sigma_*^2+\Delta_\tau.\end{aligned}$$

We can re-arrange this inequality:

$$\begin{aligned}\mathbb{E}\left[\left\|x^{T+1}-x^*\right\|^2\right]\leq&\left(1-\frac{\mu\gamma}{2}\right)^{T-\tau}\mathbb{E}\left[\left\|x^\tau-x^*\right\|^2\right]\\&+\sum_{t=\tau}^{T}13\left(1-\frac{\mu\gamma}{2}\right)^{T-t}\frac{\gamma^2d^2\tau}{m^2}\sigma_*^2+\left(1-\frac{\mu\gamma}{2}\right)^{T}\Delta_\tau.\end{aligned}$$

Using the fact that

$$\sum_{t=\tau}^{T}\left(1-\frac{\mu\gamma}{2}\right)^{T-t}=\sum_{t=0}^{T-\tau}\left(1-\frac{\mu\gamma}{2}\right)^{t}\leq\sum_{t=0}^{+\infty}\left(1-\frac{\mu\gamma}{2}\right)^{t}=\frac{2}{\mu\gamma}.$$

We can estimate:

$$\mathbb{E}\left[\left\|x^{T+1}-x^*\right\|^2\right] \leq \left(1-\frac{\mu\gamma}{2}\right)^{T-\tau}\mathbb{E}\left[\left\|x^{\tau}-x^*\right\|^2\right] + \left(1-\frac{\mu\gamma}{2}\right)^{T}\Delta_\tau + 26\frac{\gamma d^2\tau}{\mu m^2}\sigma_*^2.$$

This finishes the proof.

□

G EXTENSIONS FOR THEOREM 5

G.1 FULL VERSION OF THEOREM 5

Theorem 16 (Convergence of AMQSGD Algorithm 2, full version)**.** *Consider Assumptions 1, 2 and 4. Let problem* (1) *be solved by Algorithm 2. Then for any $\gamma > 0, \varepsilon > 0$, $\tau > \tau_{mix}(\varepsilon)$, $T > \tau$ and β, θ, η, p satisfying*

$$\gamma \lesssim \frac{\mu^{\frac{1}{3}}m^{\frac{1}{2}}}{\tau L^{\frac{4}{3}}d^{\frac{1}{2}}}, \quad p \lesssim \frac{m^2}{\tau^2 d^2(\delta^2+1)}, \quad \varepsilon \lesssim \min\left\{\frac{m^{\frac{7}{4}}}{d^{\frac{7}{4}}\tau^{\frac{5}{4}}L(\delta^2+1)}; \frac{m^{\frac{15}{4}}}{d^{\frac{15}{4}}\tau^{\frac{13}{4}}(\delta^2+1)^2}\right\}$$

$$\beta = \sqrt{\frac{2p^2\mu\gamma}{3}}, \quad \eta = \sqrt{\frac{3}{2\mu\gamma}}, \quad \theta = \frac{p\eta^{-1}-1}{\beta p\eta^{-1}-1}$$

it holds that

$$F_{T+1} = \mathcal{O}\left(\exp\left[-(T-\tau)\sqrt{\frac{p^2\mu\gamma}{3}}\right]F_\tau + \exp\left[-T\sqrt{\frac{p^2\mu\gamma}{3}}\right]\Delta_\tau + \frac{\gamma}{\mu}\sigma^2\right).$$

Here we use notations: $F_t := \mathbb{E}[\|x^t-x^*\|^2 + \frac{3}{\mu}(f(x_f^t)-f(x^*))]$ and $\Delta_\tau \leq \frac{\sqrt{\gamma}}{\tau^{\frac{4}{3}}\mu^{\frac{1}{3}}}\sum\limits_{t=0}^{\tau}\left(\mathbb{E}\left\|\nabla f(x_g^t)\right\| + \mathbb{E}\left\|x^t-x^*\right\|^2 + \mathbb{E}[f(x_f^t)-f(x^*)]\right)$.

G.2 FULL VERSION OF COROLLARY 6

Corollary 17 (Step tuning for Theorem 5, full version of Corollary 6)**.** *Under the conditions of Theorem 5, choosing γ as*

$$\gamma \lesssim \min\left\{\frac{\mu^{\frac{1}{3}}}{L^{\frac{4}{3}}\tau^{\frac{8}{3}}}\ ;\ \frac{\log\left(\max\left\{2; \frac{\mu^{\frac{2}{3}}(F_\tau+\Delta_\tau)T}{\tau^{\frac{4}{3}}L^{\frac{2}{3}}\sigma^2}\right\}\right)}{\mu p^2T^2}\right\},$$

in order to achieve ϵ-approximate solution (in terms of $\mathbb{E}\left[\left\|x^T-x^*\right\|^2\right] \leq \epsilon^2$) it takes

$$\mathcal{O}\left(\frac{d^2L^{\frac{2}{3}}\tau^{\frac{4}{3}}}{m^2\mu^{\frac{2}{3}}}\left((\delta^2+1)\log\left(\frac{1}{\epsilon}\right)+\frac{\sigma^2}{\mu\epsilon}\right)\right) \text{ iterations.}$$

G.3 PROOF OF THEOREM 16

Lemma 18. *Consider Algorithm 2 with $\theta = (p\eta^{-1}-1)/(\beta\eta^{-1}1-1) < 1$. Then for any $y^t = \kappa x_f^t + (1-\kappa)x^t \in \text{conv}\left\{x_f^t, x^t\right\}$ for any $s < t$ exist constants $\alpha_f^s, \alpha^s \geq 0$ and $c_r \geq 0$ such that*

$$y^t = \widetilde{y}^s - p\gamma\sum_{r=s}^{t-1}c_r g^r = \alpha_f^s x_f^s + \alpha^s x^s - p\gamma\sum_{r=s}^{t-1}c_r g^r.$$

And $\alpha_f^s + \alpha^s = 1$ for any $s < t$. If $(1-\kappa)\eta \leq 1$, then $c_r \leq t-s+2$, otherwise we can only use the estimate $c_r \leq \eta$.

Proof. We start by writing out lines 3 and 10 of Algorithm 2:

$$x_f^s = x_g^{s-1} - p\gamma g^{s-1} = \theta x_f^{s-1} + (1-\theta)x^{s-1} - p\gamma g^{s-1}. \tag{28}$$

Now let us handle expression $\eta x_g^k + (p-\eta)x_f^k + (1-p)(1-\beta)x^k + (1-p)\beta x_g^k - x^*$ for a while. Taking into account the choice of θ such that $\theta = (p\eta^{-1}-1)/(\beta p\eta^{-1}-1)$ (in particular, $(p\eta^{-1}-1) = (\beta p\eta^{-1}-1)\theta$ and $\eta(1-\beta p\eta^{-1})(1-\theta) = p(1-\beta)$), we get

$$\begin{aligned}
&\eta x_g^k + (p-\eta)x_f^k + (1-p)(1-\beta)x^k + (1-p)\beta x_g^k \\
&\qquad = (\eta + (1-p)\beta)x_g^k + (p-\eta)x_f^k + (1-p)(1-\beta)x^k \\
&\qquad = (\eta + (1-p)\beta)x_g^k + \eta(p\eta^{-1}-1)x_f^k + (1-p)(1-\beta)x^k \\
&\qquad = (\eta + (1-p)\beta)x_g^k + \eta(\beta p\eta^{-1}-1)\theta x_f^k + (1-p)(1-\beta)x^k \\
&\qquad = (\eta + (1-p)\beta)x_g^k + \eta(\beta p\eta^{-1}-1)(x_g^k - (1-\theta)x^k) + (1-p)(1-\beta)x^k \\
&\qquad = (\eta + (1-p)\beta)x_g^k + \eta(\beta p\eta^{-1}-1)(x_g^k - (1-\theta)x^k) + (1-p)(1-\beta)x^k \\
&\qquad = \beta x_g^k - \eta(\beta p\eta^{-1}-1)(1-\theta)x^k + (1-p)(1-\beta)x^k \\
&\qquad = \beta x_g^k + p(1-\beta)x^k + (1-p)(1-\beta)x^k \\
&\qquad = \beta x_g^k + (1-\beta)x^k\,.
\end{aligned}$$

Now we write out line 11 of Algorithm 2:

$$\begin{aligned}
x^s &= \beta x_g^{s-1} + (1-\beta)x^{s-1} - \eta x_g^{s-1} + \eta x_f^s = \beta x_g^{s-1} + (1-\beta)x^{s-1} - \eta p\gamma g^{s-1} \\
&= \beta(\theta x_f^{s-1} + (1-\theta)x^{s-1}) + (1-\beta)x^{s-1} - \eta p\gamma g^{s-1} \\
&= \beta\theta x_f^{s-1} + (1-\beta\theta)x^{s-1} - \eta p\gamma g^{s-1}.
\end{aligned} \tag{29}$$

Now we use induction. $x_f^t = \theta x_f^{s-1} + (1-\theta)x^{s-1} - p\gamma g^{s-1}$, then $\alpha_f^{t-1} = \theta \geq 0$, $\alpha^{t-1} = 1-\theta \geq 0$, $c_r = 1 \leq \eta$ and $\alpha_f^{t-1} + \alpha^{t-1} = 1$, therefore base step is fulfilled. If $x_f^t = \alpha_f^s x_f^s + \alpha^s x^s - p\gamma\sum_{r=s}^{t-1} c_r g^r$ for some $s < t$, when with help of (28) and (29) we can write out

$$\begin{aligned}
x_f^t = {}& \alpha_f^s\left(\theta x_f^{s-1} + (1-\theta)x^{s-1} - p\gamma g^{s-1}\right) \\
&+ \alpha^s\left(\beta\theta x_f^{s-1} + (1-\beta\theta)x^{s-1} - \eta p\gamma g^{s-1}\right) - p\gamma\sum_{r=s}^{t-1} c_r g^r.
\end{aligned}$$

Therefore $\alpha_f^{s-1} = \alpha_f^s\theta + \alpha^s\beta\theta \geq 0$, $\alpha^{s-1} = \alpha_f^s(1-\theta) + \alpha^s(1-\beta\theta) \geq 0$ and $c_{s-1} = \alpha_f^s + \eta\alpha^s \leq \eta$. Then, the step of the induction is fulfilled, since $\alpha_f^{s-1} + \alpha^{s-1} = 1$. Therefore results of this Lemma are true for $y^t = x_f^t \in \text{conv}\left\{x_f^t, x^t\right\}$.

Consider $y^t = x^t \in \text{conv}\left\{x_f^t, x^t\right\}$. Form (29) follows that $\alpha_f^{t-1} = \beta\theta$ and $\alpha^{t-1} = 1-\beta\theta$, therefore base step is fulfilled. The step of the induction will be the same as in $y^t = x_f^t$. Therefore results of this Lemma are true for $y^t = x^t$. Then, they are true for any $y^t \in \text{conv}\left\{x_f^t, x^t\right\}$.

If $y^t = \kappa x_f^t + (1-\kappa)x^t$, then $\alpha^s(y) = \kappa\alpha^s(x_f^t) + (1-\kappa)\alpha^s(x^t)$. Since $(1-\theta)\eta \leq 1$, then $\alpha^{t-1}(x_f^t)\eta \leq 1 = t-(t-1)$. Therefore $\alpha^s(x_f^t)\eta \leq t-s$ by induction, since $\alpha^{s-1}(x_f^t)\eta = \alpha_f^s(x_f^t)(1-\theta)\eta + (1-\beta\theta)\alpha^s(x_f^t)\eta \leq \alpha_f^s(x_f^t) + (1-\beta\theta)(t-s) \leq t-s+1$.

Then, if $(1-\kappa)\eta \leq 1$, then $\alpha^s(y^t)\eta = \kappa\alpha^s(x_f^t)\eta + (1-\kappa)\eta\alpha^s(x^t) \leq \kappa(t-s) + \alpha^s(x^t) \leq t-s+1$. Now we consider $c_s(y^t)$. $c_s(y^t) = \alpha_f^s(y^t) + \alpha^s(y^t)\eta \leq \alpha_f^s(y^t) + t - s + 1 \leq t-s+2$.

□

Lemma 19. *Assume 1, 2 and 4. Then for iterates of Algorithm 2 with $\theta = (p\eta^{-1}-1-1)/(\beta p\eta^{-1}-1-1), \theta > 0, \eta \geq 1$, it holds that*

$$\begin{aligned}
&\mathbb{E}\left\|x^{t+1}-x^*\right\|^2 \\
&\quad\leq (1-\beta)(1+\frac{\beta}{4})\,\mathbb{E}\left\|x^t-x^*\right\|^2+\beta(1+\frac{\beta}{4})\,\mathbb{E}\left\|x_g^t-x^*\right\|^2+(\beta^2-\beta)\,\mathbb{E}\left\|x^t-x_g^t\right\|^2 \\
&\quad+10\frac{d^2}{m^2}(\delta^2+1)p^2\gamma^2\eta^2\,\mathbb{E}\left\|\nabla f(x_g^t)\right\|^2+p^2\gamma^2\eta^2\tau\Big(32\frac{\tau^2d^2L^2p^2\gamma^2}{m^2\beta}+\frac{5}{4}\Big)\sum_{r=t-\tau}^{t-1}\left\|g^r\right\|^2 \\
&\quad+3\varepsilon p\gamma\eta L\frac{d}{m}\sqrt{\delta^2+1}\mathbb{E}\left[\left\|x^{t-\tau}-x^*\right\|^2\right]+3\varepsilon p\gamma\eta L\frac{d}{m}\sqrt{\delta^2+1}\mathbb{E}\left[\left\|x_f^{t-\tau}-x^*\right\|^2\right] \\
&\quad-2\gamma\eta^2\,\mathbb{E}\left\langle\nabla f(x_g^t),x_g^t+(p\eta^{-1}-1)x_f^t-p\eta^{-1}x^*\right\rangle+2p\gamma\eta\left(\frac{\varepsilon d}{m\sqrt{\delta^2+1}L}+4p\gamma\eta\frac{d^2}{m^2}\right)\sigma^2.
\end{aligned}\tag{30}$$

Proof. Using lines 10 and 11 of Algorithm 2, we get

$$\begin{aligned}
\mathbb{E}\left\|x^{t+1}-x^*\right\|^2&=\mathbb{E}\left\|\eta x_f^{t+1}+(p-\eta)x_f^t+(1-p)(1-\beta)x^t+(1-p)\beta x_g^t-x^*\right\|^2 \\
&=\mathbb{E}\left\|\eta x_g^t-p\gamma\eta g^t+(p-\eta)x_f^t+(1-p)(1-\beta)x^t+(1-p)\beta x_g^t-x^*\right\|^2 \\
&=\mathbb{E}\left\|\eta x_g^t+(p-\eta)x_f^t+(1-p)(1-\beta)x^t+(1-p)\beta x_g^t-x^*\right\|^2+p^2\gamma^2\eta^2\,\mathbb{E}\left\|g^t\right\|^2 \\
&\quad-2p\gamma\eta\,\mathbb{E}\left\langle g^t,\eta x_g^t+(p-\eta)x_f^t+(1-p)(1-\beta)x^t+(1-p)\beta x_g^t-x^*\right\rangle \\
&=\underbrace{\mathbb{E}\left\|\eta x_g^t+(p-\eta)x_f^t+(1-p)(1-\beta)x^t+(1-p)\beta x_g^t-x^*\right\|^2}_{①}+\underbrace{p^2\gamma^2\eta^2\,\mathbb{E}\left\|g^t\right\|^2}_{②} \\
&\quad\underbrace{-2p\gamma\eta\,\mathbb{E}\left\langle g^t-\nabla f(x_g^t),\eta x_g^t+(p-\eta)x_f^t+(1-p)(1-\beta)x^t+(1-p)\beta x_g^t-x^*\right\rangle}_{③} \\
&\quad\underbrace{-2p\gamma\eta\,\mathbb{E}\left\langle\nabla f(x_g^t),\eta x_g^t+(p-\eta)x_f^t+(1-p)(1-\beta)x^t+(1-p)\beta x_g^t-x^*\right\rangle}_{④}.
\end{aligned}$$

Consider ①. From Lemma 18, we know that

$$\eta x_g^t+(p-\eta)x_f^t+(1-p)(1-\beta)x^t+(1-p)\beta x_g^t=\beta x_g^t+(1-\beta)x^t.$$

It implies

$$\begin{aligned}
&\left\|\eta x_g^t+(p-\eta)x_f^t+(1-p)(1-\beta)x^t+(1-p)\beta x_g^t-x^*\right\|^2 \\
&\quad=\left\|\beta x_g^t+(1-\beta)x^t-x^*\right\|^2 \\
&\quad=\left\|\beta(x_g^t-x^t)+x^t-x^*\right\|^2 \\
&\quad=\left\|x^t-x^*\right\|^2+2\beta\left\langle x^t-x^*,x_g^t-x^t\right\rangle+\beta^2\left\|x_g^t-x^t\right\|^2 \\
&\quad=\left\|x^t-x^*\right\|^2+\beta(\left\|x_g^t-x^*\right\|^2-\left\|x^t-x^*\right\|^2-\left\|x_g^t-x^t\right\|^2)+\beta^2\left\|x_g^t-x^t\right\|^2 \\
&\quad=(1-\beta)\left\|x^t-x^*\right\|^2+\beta\left\|x_g^t-x^*\right\|^2+(\beta^2-\beta)\left\|x^t-x_g^t\right\|^2.
\end{aligned}\tag{31}$$

Consider ②. Using convexity of squared Euclidean norm and Lemma 10, one can obtain

$$\begin{aligned} p^2\gamma^2\eta^2\,\mathbb{E}\left\|g^t\right\|^2 &= p^2\gamma^2\eta^2\,\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n Q_t^i(\nabla f_i(x_g^t))\right\|^2 \\ &\le p^2\gamma^2\eta^2\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left\|Q_t^i(\nabla f_i(x_g^t))\right\|^2 \\ &\overset{(10)}{\le} p^2\gamma^2\eta^2\frac{d^2}{m^2}\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left\|\nabla f_i(x_g^t)\right\|^2 \\ &\overset{(12)}{\le} 2p^2\gamma^2\eta^2\frac{d^2}{m^2}(\delta^2+1)\,\mathbb{E}\left\|\nabla f(x_g^t)\right\|^2 + 2p^2\gamma^2\eta^2\frac{d^2}{m^2}\sigma^2, \end{aligned} \tag{32}$$

where in the last inequality we used Lemma 12.
Consider ③. We first use Lemma 18 twice

$$x_g^t = \theta x_f^t + (1-\theta)x^t = \alpha_f^{t-\tau}x_f^{t-\tau} + \alpha^{t-\tau}x^{t-\tau} - p\gamma\sum_{r=t-\tau}^{t-1} c_r g^r$$

$$\begin{aligned} \eta x_g^t + (p-\eta)x_f^t + (1-p)(1-\beta)x^t + (1-p)\beta x_g^t &= \beta x_g^t + (1-\beta)x^t \\ &= \beta\theta x_f^t + (1-\beta\theta)x^t \\ &= \hat{\alpha}_f^{t-\tau}x_f^{t-\tau} + \hat{\alpha}^{t-\tau}x^{t-\tau} - p\gamma\sum_{r=t-\tau}^{t-1}\hat{c}_r g^r. \end{aligned}$$

Next, we apply Corollary 11 with $\hat{a}^{t-\tau} = \nabla f_i(\widetilde{x}_g^{t-\tau})$, where $\widetilde{x}_g^{t-\tau} = \alpha_f^{t-\tau}x_f^{t-\tau} + \alpha^{t-\tau}x^{t-\tau}$, and $\hat{b}^{t-\tau} = \hat{\alpha}_f^{t-\tau}x_f^{t-\tau} + \hat{\alpha}^{t-\tau}x^{t-\tau} - x^*$, leading us to

$$\begin{aligned} &-2p\gamma\eta\,\mathbb{E}\left\langle g^t - \nabla f(x_g^t), \eta x_g^t + (p-\eta)x_f^t + (1-p)(1-\beta)x^t + (1-p)\beta x_g^t - x^*\right\rangle \\ &= -2p\gamma\eta\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left\langle Q_t^i(\nabla f_i(x_g^t)) - \nabla f_i(x_g^t), \eta x_g^t + (p-\eta)x_f^t + (1-p)(1-\beta)x^t + (1-p)\beta x_g^t - x^*\right\rangle \\ &\le \frac{\varepsilon d}{m\beta_0}p\gamma\eta\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\nabla f_i(\widetilde{x}_g^{t-\tau})\right\|^2\right] + \frac{\varepsilon d\beta_0}{m}p\gamma\eta\mathbb{E}\left[\left\|\hat{\alpha}_f^{t-\tau}x_f^{t-\tau} + \hat{\alpha}^{t-\tau}x^{t-\tau} - x^*\right\|^2\right] \\ &\quad + 4\frac{d^2}{m^2}p\gamma\eta\,(\beta_1+\beta_2)\,\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\nabla f_i(x_g^t) - \nabla f_i(\widetilde{x}_g^{t-\tau})\right\|^2\right] + p\gamma\eta\left(\frac{1}{\beta_1}+\frac{1}{\beta_3}\right)\mathbb{E}\left[\left\|-p\gamma\sum_{r=t-\tau}^{t-1}\hat{c}_r g^r\right\|^2\right] \\ &\quad + 4\frac{d^2}{m^2}p\gamma\eta\beta_3\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\nabla f_i(x_g^t)\right\|^2\right] + \frac{p\gamma\eta}{\beta_2}\mathbb{E}\left[\left\|\beta x_g^t + (1-\beta)x^t - x^*\right\|^2\right]. \end{aligned}$$

Using Assumption 1 and Lemma 12 with $c_r \le \tau \le 2\tau$ and $\hat{c}_r \le \eta$ one might obtain

$$\begin{aligned} &-2p\gamma\eta\,\mathbb{E}\left\langle g^t - \nabla f(x_g^t), \eta x_g^t + (p-\eta)x_f^t + (1-p)(1-\beta)x^t + (1-p)\beta x_g^t - x^*\right\rangle \\ &\le \frac{2\varepsilon d}{m\beta_0}p\gamma\eta(\delta^2+1)\mathbb{E}\left[\left\|\nabla f(\widetilde{x}_g^{t-\tau})\right\|^2\right] + \frac{\varepsilon d\beta_0}{m}p\gamma\eta\mathbb{E}\left[\left\|\hat{\alpha}_f^{t-\tau}x_f^{t-\tau} + \hat{\alpha}^{t-\tau}x^{t-\tau} - x^*\right\|^2\right] \\ &\quad + 4\frac{d^2L^2}{m^2}p\gamma\eta\,(\beta_1+\beta_2)\,\mathbb{E}\left[\left\|-p\gamma\sum_{r=t-\tau}^{t-1}c_r g^r\right\|^2\right] + p\gamma\eta\left(\frac{1}{\beta_1}+\frac{1}{\beta_3}\right)\mathbb{E}\left[\left\|-p\gamma\sum_{r=t-\tau}^{t-1}\hat{c}_r g^r\right\|^2\right] \\ &\quad + 8\frac{d^2}{m^2}(\delta^2+1)p\gamma\eta\beta_3\mathbb{E}\left[\left\|\nabla f(x_g^t)\right\|^2\right] + \frac{p\gamma\eta}{\beta_2}\mathbb{E}\left[\left\|\beta x_g^t + (1-\beta)x^t - x^*\right\|^2\right] \\ &\quad + 2p\gamma\eta(\frac{\varepsilon d}{m\beta_0} + 4\frac{d^2\beta_3}{m^2})\sigma^2 \end{aligned} \tag{33}$$

$$
\begin{aligned}
&\leq \frac{\varepsilon d}{m} p\gamma\eta\left(2(\delta^2+1)L^2\alpha_f^{t-\tau}\frac{1}{\beta_0}+\beta_0\hat{\alpha}_f^{t-\tau}\right)\mathbb{E}\left[\left\|x_f^{t-\tau}-x^*\right\|^2\right]\\
&+\frac{\varepsilon d}{m} p\gamma\eta\left(2(\delta^2+1)L^2\alpha^{t-\tau}\frac{1}{\beta_0}+\beta_0\hat{\alpha}^{t-\tau}\right)\mathbb{E}\left[\left\|x^{t-\tau}-x^*\right\|^2\right]\\
&+p^3\gamma^3\eta\tau\left(4\frac{\tau^2d^2L^2}{m^2}(\beta_1+\beta_2)+\eta^2\Big(\frac{1}{\beta_1}+\frac{1}{\beta_3}\Big)\right)\sum_{r=t-\tau}^{t-1}\left\|g^r\right\|^2\\
&+8\frac{d^2}{m^2}(\delta^2+1)p\gamma\eta\beta_3\mathbb{E}\left[\left\|\nabla f(x_g^t)\right\|^2\right]\\
&+\frac{p\gamma\eta}{\beta_2}\beta\mathbb{E}\left[\left\|x_g^t-x^*\right\|^2\right]+\frac{p\gamma\eta}{\beta_2}(1-\beta)\mathbb{E}\left[\left\|x^t-x^*\right\|^2\right]+2p\gamma\eta(\frac{\varepsilon d}{m\beta_0}+4\frac{d^2\beta_3}{m^2})\sigma^2.
\end{aligned}
$$

Consider ④. Taking into account line 4 and the choice of θ such that $\theta=(p\eta^{-1}-1)/(\beta p\eta^{-1}-1)$, one can note

$$
\begin{aligned}
&\eta x_g^k+(p-\eta)x_f^k+(1-p)(1-\beta)x^k+(1-p)\beta x_g^k-x^*\\
&=(\eta+(1-p)\beta)x_g^k+(p-\eta)x_f^k+(1-p)(1-\beta)x^k-x^*\\
&=\eta p^{-1}\left((p+(1-p)p^{-1}\eta\beta)x_g^k+(p\eta^{-1}-1)px_f^k+(1-p)(1-\beta)p\eta^{-1}x^k-\eta^{-1}px^*\right)\\
&=\eta p^{-1}\left((p+(1-p)p^{-1}\eta\beta)x_g^k+(p\eta^{-1}-1)px_f^k+(1-p)(1-\beta p\eta^{-1})(1-\theta)x^k-\eta^{-1}px^*\right)\\
&=\eta p^{-1}\left((p+(1-p)p^{-1}\eta\beta)x_g^k+(p\eta^{-1}-1)px_f^k+(1-p)(1-\beta p\eta^{-1})(x_g^k-\theta x_f^k)-\eta^{-1}px^*\right)\\
&=\eta p^{-1}\left(x_g^k+(p\eta^{-1}-1)px_f^k-(1-p)(1-\beta p\eta^{-1})\theta x_f^k-\eta^{-1}px^*\right)\\
&=\eta p^{-1}\left(x_g^k+(p\eta^{-1}-1)px_f^k-(1-p)(p\eta^{-1}-1)x_f^k-\eta^{-1}px^*\right)\\
&=\eta p^{-1}\left(x_g^k+(p\eta^{-1}-1)x_f^k-\eta^{-1}px^*\right).
\end{aligned}
\tag{34}
$$

Using that, we get

$$
\begin{aligned}
&-2p\gamma\eta\,\mathbb{E}\left\langle\nabla f(x_g^t),\eta x_g^t+(p-\eta)x_f^t+(1-p)(1-\beta)x^t+(1-p)\beta x_g^t-x^*\right\rangle\\
&\qquad=-2\gamma\eta^2\,\mathbb{E}\left\langle\nabla f(x_g^t),x_g^t+(p\eta^{-1}-1)x_f^t-p\eta^{-1}x^*\right\rangle.
\end{aligned}
\tag{35}
$$

Summing (31), (32), (33) and (35) with $\beta_0=\sqrt{\delta^2+1}L,\ \beta_1=\beta_2=\frac{4p\gamma\eta}{\beta}$ and $\beta_3=p\gamma\eta$ we finish the proof. □

Lemma 20. *Assume 1, 2 and 4. Then for iterates of Algorithm 2 and for any $u\in\mathbb{R}^d$ it holds that*

$$
\begin{aligned}
&\mathbb{E}\left[f(x_f^{t+1})\right]\leq\mathbb{E}\left[f(u)\right]-\mathbb{E}\left[\left\langle\nabla f(x_g^t),u-x_g^t\right\rangle\right]-\frac{\mu}{2}\left\|u-x_g^t\right\|-\frac{p\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x_g^t)\right\|^2\right]\\
&+2\varepsilon\gamma\mathbb{E}\left[\left\|\nabla f(\widetilde{x}_g^{t-\tau})\right\|^2\right]+20\frac{L^2d^3\gamma^3p^2\tau^3(\delta^2+1)}{m^3}\sum_{s=t-\tau}^{t-1}\mathbb{E}\left[\left\|\nabla f(x_g^s)\right\|^2\right]+23\frac{L^2d^3\gamma^3p^2\tau^4}{m^3}\sigma^2,
\end{aligned}
$$

where

$$
\gamma\leq\frac{1}{L}\quad\textit{and}\quad p\leq\frac{m^2}{12(\delta^2+1)d^2}.
$$

Proof. Using 1 with $x=x_f^{t+1}$, $y=x_g^t$ and line 3 of Algorithm 2 we get

$$
\begin{aligned}
\mathbb{E}\left[f(x_f^{t+1})\right]&\leq\mathbb{E}\left[f(x_g^t)\right]+\mathbb{E}\left[\left\langle\nabla f(x_g^t),x_f^{t+1}-x_g^t\right\rangle\right]+\frac{L}{2}\mathbb{E}\left[\left\|x_f^{t+1}-x_g^t\right\|^2\right]\\
&=\mathbb{E}\left[f(x_g^t)\right]-p\gamma\mathbb{E}\left[\left\langle\nabla f(x_g^t),g^t\right\rangle\right]+\frac{Lp^2\gamma^2}{2}\mathbb{E}\left[\left\|g^t\right\|^2\right]\\
&=\mathbb{E}\left[f(x_g^t)\right]-p\gamma\mathbb{E}\left[\left\langle\nabla f(x_g^t),\nabla f(x_g^t)\right\rangle\right]-p\gamma\mathbb{E}\left[\left\langle\nabla f(x_g^t),g^k-\nabla f(x_g^t)\right\rangle\right]\\
&\quad+\frac{Lp^2\gamma^2}{2}\mathbb{E}\left[\left\|g^t\right\|^2\right].
\end{aligned}
\tag{36}
$$

Consider $\mathbb{E}\left[\left\langle \nabla f(x_g^t), g^k - \nabla f(x_g^t)\right\rangle\right]$. Using Corollary 11 with $a^t = \nabla f_i(x_g^t), b^t = \nabla f(x_g^t), \hat{a}^{t-\tau} = \nabla f_i(\widetilde{x}_g^{t-\tau}), \hat{b}^{t-\tau} = \nabla f(\widetilde{x}_g^{t-\tau})$, where $x_g^t \in \text{conv}\left\{x_f^t, x^t\right\} = \widetilde{x}_g^{t-\tau} - p\gamma\sum_{s=t-\tau}^{t-1} c_s g^s$ from Lemma 18. Using Assumption 1 we obtain

$$
\begin{aligned}
2\left|\mathbb{E}\left[\left\langle \nabla f(x_g^t), g^k - \nabla f(x_g^t)\right\rangle\right]\right| &\leq \frac{\varepsilon d}{m\beta_0}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(\widetilde{x}_g^{t-\tau})\right\|^2\right] + \frac{\varepsilon d\beta_0}{m}\mathbb{E}\left[\left\|\nabla f(\widetilde{x}_g^{t-\tau})\right\|^2\right] \\
&+ 4\frac{d^2L^2}{m^2}(\beta_1+\beta_2)\mathbb{E}\left[\left\|x_g^t - \widetilde{x}_g^{t-\tau}\right\|^2\right] + L^2\left(\frac{1}{\beta_1}+\frac{1}{\beta_3}\right)\mathbb{E}\left[\left\|x_g^t - \widetilde{x}_g^{t-\tau}\right\|^2\right] \\
&+ 4\frac{d^2}{m^2}\beta_3\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(x_g^t)\right\|^2\right] + \frac{1}{\beta_2}\mathbb{E}\left[\left\|\nabla f(x_g^t)\right\|^2\right].
\end{aligned}
$$

Taking $\beta_0 = \sqrt{\delta^2+1}$, $\beta_1 = m/d$, $\beta_2 = m/(dp)$, $\beta_3 = pm/d$ and using results from Lemma 12 we obtain

$$
\begin{aligned}
2\left|\mathbb{E}\left[\left\langle \nabla f(x_g^t), g^k - \nabla f(x_g^t)\right\rangle\right]\right| &\leq \frac{2\varepsilon d}{m}\left(\sqrt{\delta^2+1}\mathbb{E}\left[\left\|\nabla f(\widetilde{x}_g^{t-\tau})\right\|^2\right] + \frac{\sigma^2}{\sqrt{\delta^2+1}}\right) \\
&+ \frac{dp}{m}\mathbb{E}\left[\left\|\nabla f(x_g^t)\right\|^2\right] + 10\frac{L^2 d}{pm}\mathbb{E}\left[\left\|-p\gamma\sum_{s=t-\tau}^{t-1} c_s \frac{1}{n}\sum_{i=1}^n Q_s^i(\nabla f_i(x_g^s))\right\|^2\right] \\
&+ \frac{8dp}{m}\left((\delta^2+1)\mathbb{E}\left[\left\|\nabla f(x_g^t)\right\|^2\right] + \sigma^2\right) + \frac{\varepsilon d\sqrt{\delta^2+1}}{m}\mathbb{E}\left[\left\|\nabla f(\widetilde{x}_g^{t-\tau})\right\|^2\right].
\end{aligned}
$$

Using Lemma 10 and 12, convexity of the squared norm and the fact that $c_s \leq t - s + 2 \leq \tau + 2 \leq 2\tau$ we obtain

$$
\begin{aligned}
2\left|\mathbb{E}\left[\left\langle \nabla f(x_g^t), g^k - \nabla f(x_g^t)\right\rangle\right]\right| &\leq \frac{3\varepsilon d\sqrt{\delta^2+1}}{m}\mathbb{E}\left[\left\|\nabla f(\widetilde{x}_g^{t-\tau})\right\|^2\right] + \\
&+ 40\frac{L^2 d^3\gamma^2 p\tau^3}{m^3}\sum_{s=t-\tau}^{t-1}\mathbb{E}\left[(\delta^2+1)\left\|\nabla f(x_g^s)\right\|^2 + \sigma^2\right] \\
&+ \frac{9dp(\delta^2+1)}{m}\mathbb{E}\left[\left\|\nabla f(x_g^t)\right\|^2\right] + \frac{2d}{m}\left(\frac{\varepsilon}{\sqrt{\delta^2+1}} + p\right)\sigma^2.
\end{aligned}
$$

Using the fact that $L^2\gamma^2 d^2/m^2\tau^4\eta^2 \geq 1$ and $\varepsilon \leq \sqrt{\delta^2+1}p$ we obtain

$$
\begin{aligned}
2\left|\mathbb{E}\left[\left\langle \nabla f(x_g^t), g^k - \nabla f(x_g^t)\right\rangle\right]\right| &\leq \frac{3\varepsilon d\sqrt{\delta^2+1}}{m}\mathbb{E}\left[\left\|\nabla f(\widetilde{x}_g^{t-\tau})\right\|^2\right] + 44\frac{L^2 d^3\gamma^2 p\eta^2\tau^4}{m^3}\sigma^2 \\
&+ 40\frac{L^2 d^3\gamma^2 p\tau^3(\delta^2+1)}{m^3}\sum_{s=t-\tau}^{t-1}\mathbb{E}\left[\left\|\nabla f(x_g^s)\right\|^2\right] + \frac{9dp(\delta^2+1)}{m}\mathbb{E}\left[\left\|\nabla f(x_g^t)\right\|^2\right].
\end{aligned}
$$

Using this result, Lemmas 10 and 12 we can estimate (36):

$$
\begin{aligned}
\mathbb{E}\left[f(x_f^{t+1})\right] &= \mathbb{E}\left[f(x_g^t)\right] - p\gamma\mathbb{E}\left[\left\|\nabla f(x_g^t)\right\|^2\right] \\
&- p\gamma\mathbb{E}\left[\left\langle \nabla f(x_g^t), g^k - \nabla f(x_g^t)\right\rangle\right] + \frac{L}{2}\mathbb{E}\left[\left\|g^t\right\|^2\right] \\
&\leq \mathbb{E}\left[f(x_g^t)\right] - p\gamma\mathbb{E}\left[\left\|\nabla f(x_g^t)\right\|^2\right] + \frac{2\varepsilon p\gamma d\sqrt{\delta^2+1}}{m}\mathbb{E}\left[\left\|\nabla f(\widetilde{x}_g^{t-\tau})\right\|^2\right] + \\
&+ 20\frac{L^2 d^3\gamma^3 p^2\tau^3(\delta^2+1)}{m^3}\sum_{s=t-\tau}^{t-1}\mathbb{E}\left[\left\|\nabla f(x_g^s)\right\|^2\right] + \frac{5d\gamma p^2(\delta^2+1)}{m}\mathbb{E}\left[\left\|\nabla f(x_g^t)\right\|^2\right]
\end{aligned}
$$

$$+ 22\frac{L^2d^3\gamma^3p^2\tau^4}{m^3}\sigma^2 + \frac{Lp^2\gamma^2d^2}{m^2}(\delta^2+1)\mathbb{E}\left[\left\|\nabla f(x_g^t)\right\|^2\right] + \frac{Lp^2\gamma^2d^2}{m^2}\sigma^2.$$

Taking

$$\gamma \leq \frac{1}{L} \quad \text{and} \quad p \leq \frac{m^2}{12(\delta^2+1)d^2},$$

we obtain

$$\mathbb{E}\left[f(x_f^{t+1})\right] \leq \mathbb{E}\left[f(x_g^t)\right] - \frac{p\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x_g^t)\right\|^2\right] + 2\varepsilon\gamma\mathbb{E}\left[\left\|\nabla f(\widetilde{x}_g^{t-\tau})\right\|^2\right] + \\ + 20\frac{L^2d^3\gamma^3p^2\tau^3(\delta^2+1)}{m^3}\sum_{s=t-\tau}^{t-1}\mathbb{E}\left[\left\|\nabla f(x_g^s)\right\|^2\right] + 23\frac{L^2d^3\gamma^3p^2\tau^4}{m^3}\sigma^2.$$

Using 2 with $x = u$ and $y = x_g^t$, one can conclude that for any $u \in \mathbb{R}^d$ it holds

$$\mathbb{E}\left[f(x_f^{t+1})\right] \leq \mathbb{E}\left[f(u)\right] - \mathbb{E}\left[\left\langle\nabla f(x_g^t), u - x_g^t\right\rangle\right] - \frac{\mu}{2}\left\|u - x_g^t\right\| \\ - \frac{p\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x_g^t)\right\|^2\right] + 2\varepsilon\gamma\mathbb{E}\left[\left\|\nabla f(\widetilde{x}_g^{t-\tau})\right\|^2\right] + \\ + 20\frac{L^2d^3\gamma^3p^2\tau^3(\delta^2+1)}{m^3}\sum_{s=t-\tau}^{t-1}\mathbb{E}\left[\left\|\nabla f(x_g^s)\right\|^2\right] + 23\frac{L^2d^3\gamma^3p^2\tau^4}{m^3}\sigma^2.$$

This finishes the proof. □

Theorem 21 (Theorem 5)**.** *Consider Assumptions 1, 2 and 4. Let problem* (1) *be solved by Algorithm 2. Then for any* $\gamma > 0, \varepsilon > 0$*,* $\tau > \tau_{mix}(\varepsilon)$*,* $T > \tau$ *and* β, θ, η, p *satisfying*

$$\gamma \leq \frac{\mu^{\frac{1}{3}}m^{\frac{1}{2}}}{2\tau L^{\frac{4}{3}}d^{\frac{1}{2}}}, \quad \varepsilon \leq \min\left\{\frac{m^{\frac{7}{4}}}{6d^{\frac{7}{4}}\tau^{\frac{5}{4}}L(\delta^2+1)}; \frac{m^{\frac{5}{4}}}{\sqrt{2}\tau^{\frac{3}{4}}\mu^{\frac{1}{3}}L^{\frac{2}{3}}d^{\frac{5}{4}}}; \frac{m^{\frac{15}{4}}}{6d^{\frac{15}{4}}\tau^{\frac{13}{4}}(\delta^2+1)^2}\right\},$$

$$p \leq \frac{m^2}{13d^2(\delta^2+1)\tau^2}, \quad \beta = \sqrt{\frac{2p^2\mu\gamma}{3}}, \quad \eta = \sqrt{\frac{3}{2\mu\gamma}}, \quad \theta = \frac{p\eta^{-1}-1}{\beta p\eta^{-1}-1}.$$

it holds that

$$\mathbb{E}[\|x^{T+1} - x^*\|^2 + \frac{3}{\mu}(f(x_f^{T+1}) - f(x^*))] \leq \exp\left(-(T-\tau)\sqrt{\frac{2p^2\mu\gamma}{3}}\right)F_\tau \\ + \exp\left(-T\sqrt{\frac{2p^2\mu\gamma}{3}}\right)\Delta_\tau + \frac{45\gamma}{\mu}\sigma^2,$$

where $F_\tau := \mathbb{E}[\|x^\tau - x^*\|^2 + \frac{3}{\mu}(f(x_f^\tau) - f(x^*))]$ *and* $\Delta_\tau \leq \frac{\sqrt{\gamma}}{\tau^{\frac{4}{3}}\mu^{\frac{1}{3}}}\sum_{t=0}^{\tau}\left(\mathbb{E}\left\|\nabla f(x_g^t)\right\| + \mathbb{E}\left\|x^t - x^*\right\|^2 + \mathbb{E}[f(x_f^t) - f(x^*)]\right)$.

Proof. We start by using Lemma 20 with $u = x^*$ and $u = x_f^t$

$$\mathbb{E}\left[f(x_f^{t+1})\right] \leq \mathbb{E}\left[f(x^*)\right] - \mathbb{E}\left[\left\langle\nabla f(x_g^t), x^* - x_g^t\right\rangle\right] - \frac{\mu}{2}\left\|x^* - x_g^t\right\| - \frac{p\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x_g^t)\right\|^2\right] \\ + 2\varepsilon\gamma\mathbb{E}\left[\left\|\nabla f(\widetilde{x}_g^{t-\tau})\right\|^2\right] + 20\frac{L^2d^3\gamma^3p^2\tau^3(\delta^2+1)}{m^3}\sum_{s=t-\tau}^{t-1}\mathbb{E}\left[\left\|\nabla f(x_g^s)\right\|^2\right] + 23\frac{L^2d^3\gamma^3p^2\tau^4}{m^3}\sigma^2,$$

$$\mathbb{E}\left[f(x_f^{t+1})\right] \le \mathbb{E}\left[f(x_f^t)\right] - \mathbb{E}\left[\langle \nabla f(x_g^t), x_f^t - x_g^t\rangle\right] - \frac{\mu}{2}\left\|x_f^t - x_g^t\right\| - \frac{p\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x_g^t)\right\|^2\right]$$
$$+ 2\varepsilon\gamma\mathbb{E}\left[\left\|\nabla f(\widetilde{x}_g^{t-\tau})\right\|^2\right] + 20\frac{L^2d^3\gamma^3p^2\tau^3(\delta^2+1)}{m^3}\sum_{s=t-\tau}^{t-1}\mathbb{E}\left[\left\|\nabla f(x_g^s)\right\|^2\right] + 23\frac{L^2d^3\gamma^3p^2\tau^4}{m^3}\sigma^2.$$

Summing the first inequality with coefficient $2p\gamma\eta$, the second with coefficient $2p\gamma\eta(\eta-p)$ and (30), we get

$$\begin{aligned}
&\mathbb{E}[\|x^{t+1}-x^*\|^2 + 2\gamma\eta^2 f(x_f^{t+1})] \\
&\quad\le (1-\beta)(1+\frac{\beta}{4})\,\mathbb{E}\left\|x^t - x^*\right\|^2 + \beta(1+\frac{\beta}{4})\,\mathbb{E}\left\|x_g^t - x^*\right\|^2 + (\beta^2-\beta)\,\mathbb{E}\left\|x^t - x_g^t\right\|^2 \\
&\qquad + 10\frac{d^2}{m^2}(\delta^2+1)p^2\gamma^2\eta^2\,\mathbb{E}\left\|\nabla f(x_g^t)\right\|^2 + p^2\gamma^2\eta^2\tau\Big(32\frac{\tau^2d^2L^2p^2\gamma^2}{m^2\beta} + \frac{5}{4}\Big)\sum_{r=t-\tau}^{t-1}\|g^r\|^2 \\
&\qquad + 3\varepsilon p\gamma\eta L\frac{d}{m}\sqrt{\delta^2+1}\mathbb{E}\left[\left\|x^{t-\tau}-x^*\right\|^2\right] + 3\varepsilon p\gamma\eta L\frac{d}{m}\sqrt{\delta^2+1}\mathbb{E}\left[\left\|x_f^{t-\tau}-x^*\right\|^2\right] \\
&\qquad - 2\gamma\eta^2\,\mathbb{E}\left\langle\nabla f(x_g^t), x_g^t + (p\eta^{-1}-1)x_f^t - p\eta^{-1}x^*\right\rangle + 2p\gamma\eta\left(\frac{\varepsilon d}{m\sqrt{\delta^2+1}L} + 4p\gamma\eta\frac{d^2}{m^2}\right)\sigma^2 \\
&\qquad + 2p\gamma\eta\Bigg(\mathbb{E}\left[f(x^*)\right] - \mathbb{E}\left[\langle\nabla f(x_g^t), x^* - x_g^t\rangle\right] - \frac{\mu}{2}\left\|x^* - x_g^t\right\| - \frac{p\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x_g^t)\right\|^2\right] \\
&\qquad + 2\varepsilon\gamma\mathbb{E}\left[\left\|\nabla f(\widetilde{x}_g^{t-\tau})\right\|^2\right] + 20\frac{L^2d^3\gamma^3p^2\tau^3(\delta^2+1)}{m^3}\sum_{s=t-\tau}^{t-1}\mathbb{E}\left[\left\|\nabla f(x_g^s)\right\|^2\right] \\
&\qquad + 23\frac{L^2d^3\gamma^3p^2\tau^4}{m^3}\sigma^2\Bigg) \\
&\qquad + 2\gamma\eta(\eta-p)\Bigg(\mathbb{E}\left[f(x_f^t)\right] - \mathbb{E}\left[\langle\nabla f(x_g^t), x_f^t - x_g^t\rangle\right] - \frac{\mu}{2}\left\|x_f^t - x_g^t\right\| - \frac{p\gamma}{2}\mathbb{E}\left[\left\|\nabla f(x_g^t)\right\|^2\right] \\
&\qquad + 2\varepsilon\gamma\mathbb{E}\left[\left\|\nabla f(\widetilde{x}_g^{t-\tau})\right\|^2\right] + 20\frac{L^2d^3\gamma^3p^2\tau^3(\delta^2+1)}{m^3}\sum_{s=t-\tau}^{t-1}\mathbb{E}\left[\left\|\nabla f(x_g^s)\right\|^2\right] \\
&\qquad + 23\frac{L^2d^3\gamma^3p^2\tau^4}{m^3}\sigma^2\Bigg) \\
&\quad\le (1-\beta)(1+\frac{\beta}{4})\,\mathbb{E}\left\|x^t - x^*\right\|^2 + (\beta + \frac{\beta^2}{4} - p\gamma\eta\mu)\,\mathbb{E}\left\|x_g^t - x^*\right\|^2 + (\beta^2-\beta)\,\mathbb{E}\left\|x^t - x_g^t\right\|^2 \\
&\qquad + p^2\gamma^2\eta^2\left(10\frac{d^2}{m^2}(\delta^2+1) - \frac{1}{p}\right)\mathbb{E}\left\|\nabla f(x_g^t)\right\| + 2p\gamma\eta\,\mathbb{E}\,f(x^*) + 2\gamma\eta(\eta-p)\,\mathbb{E}\,f(x_f^t) \\
&\qquad + p^2\gamma^2\eta^2\tau(\delta^2+1)\frac{d^2}{m^2}\Big(32\frac{\tau^2d^2L^2p^2\gamma^2}{m^2\beta} + \frac{5}{4}\Big)\sum_{r=t-\tau}^{t-1}\mathbb{E}\left\|\nabla f(x_g^r)\right\| \\
&\qquad + \varepsilon\gamma\eta L(3p\frac{d}{m}\sqrt{\delta^2+1} + 2\gamma\eta L)\mathbb{E}\left[\left\|x^{t-\tau}-x^*\right\|^2\right] \\
&\qquad + \varepsilon\gamma\eta L(3p\frac{d}{m}\sqrt{\delta^2+1} + 2\gamma\eta L)\mathbb{E}\left[\left\|x_f^{t-\tau}-x^*\right\|^2\right] \\
&\qquad + 2p\gamma\eta\Bigg(\frac{\varepsilon d}{m\sqrt{\delta^2+1}L} + 4p\gamma\eta\frac{d^2}{m^2} \\
&\qquad\qquad + 23p\gamma^3\eta\tau^4\frac{d^3}{m^3}L^2 + p\gamma\eta\tau^2\frac{d^2}{m^2}\left(16\frac{\tau^2d^2L^2p^2\gamma^2}{m^2\beta} + \frac{5}{8}\right)\Bigg)\sigma^2,
\end{aligned}$$

where in the last inequality we used Lemma 12 and Assumption 1. Since $\beta < 1$, the choice of $p\gamma\eta\mu = \frac{3\beta}{2}$ gives

$$
\begin{aligned}
&(1-\beta)(1+\frac{\beta}{4}) \leq 1 - \frac{3\beta}{4},\\
&\beta + \frac{\beta^2}{4} - p\gamma\eta\mu \leq \frac{3\beta}{2} - p\gamma\eta\mu \leq 0,\\
&\beta^2 - \beta \leq 0.
\end{aligned}
$$

This lead us to

$$
\begin{aligned}
\mathbb{E}[\|x^{t+1} - x^*\|^2 + 2\gamma\eta^2(f(x_f^{t+1}) - f(x^*))] &\leq (1 - \frac{3\beta}{4})\,\mathbb{E}\left\|x^t - x^*\right\|^2 + 2p\gamma\eta^2(1-\frac{p}{\eta})\,\mathbb{E}[f(x_f^t) - f(x^*)]\\
&+ p^2\gamma^2\eta^2\left(10\frac{d^2}{m^2}(\delta^2+1) - \frac{1}{p}\right)\mathbb{E}\left\|\nabla f(x_g^t)\right\| + p^2\gamma^2\eta^2\tau(\delta^2+1)\frac{d^2}{m^2}\Big(32\frac{\tau^2d^2L^2p^2\gamma^2}{m^2\beta} + \frac{5}{4}\Big)\sum_{r=t-\tau}^{t-1}\mathbb{E}\left\|\nabla f(x_g^r)\right\|\\
&+ \varepsilon\gamma\eta L(3p\frac{d}{m}\sqrt{\delta^2+1} + 2\gamma\eta L)\mathbb{E}\left[\left\|x^{t-\tau} - x^*\right\|^2\right] + \varepsilon\gamma\eta L(3p\frac{d}{m}\sqrt{\delta^2+1} + 2\gamma\eta L)\frac{2}{\mu}\,\mathbb{E}[f(x_f^{t-\tau}) - f(x^*)]\\
&+ 2p\gamma\eta\left(\frac{\varepsilon d}{m\sqrt{\delta^2+1}L} + 4p\gamma\eta\frac{d^2}{m^2} + 23p\gamma^3\eta\tau^4\frac{d^3}{m^3}L^2 + p\gamma\eta\tau^2\frac{d^2}{m^2}\left(16\frac{\tau^2d^2L^2p^2\gamma^2}{m^2\beta} + \frac{5}{8}\right)\right)\sigma^2,
\end{aligned}
\tag{37}
$$

where we also used Assumption 2 and subtracted $2\gamma\eta^2 f(x^*)$ from both sides. Next, we perform the summation from $t = \tau$ to $t = T > \tau$ of equations (37) with coefficients p_t:

$$
\begin{aligned}
&\sum_{t=\tau}^{T} p_t\,\mathbb{E}[\|x^{t+1} - x^*\|^2 + 2\gamma\eta^2(f(x_f^{t+1}) - f(x^*))]\\
&\leq \sum_{t=\tau}^{T} p_t(1 - \frac{3\beta}{4})\,\mathbb{E}\left\|x^t - x^*\right\|^2\\
&+ \sum_{t=\tau}^{T} p_t 2p\gamma\eta^2(1-\frac{p}{\eta})\,\mathbb{E}[f(x_f^t) - f(x^*)] + \sum_{t=\tau}^{T} p_t p^2\gamma^2\eta^2\left(10\frac{d^2}{m^2}(\delta^2+1) - \frac{1}{p}\right)\mathbb{E}\left\|\nabla f(x_g^t)\right\|\\
&+ \sum_{t=\tau}^{T} p_t p^2\gamma^2\eta^2\tau(\delta^2+1)\frac{d^2}{m^2}\Big(32\frac{\tau^2d^2L^2p^2\gamma^2}{m^2\beta} + \frac{5}{4}\Big)\sum_{r=t-\tau}^{t-1}\mathbb{E}\left\|\nabla f(x_g^r)\right\|\\
&+ \sum_{t=\tau}^{T} p_t\varepsilon\gamma\eta L(3p\frac{d}{m}\sqrt{\delta^2+1} + 2\gamma\eta L)\mathbb{E}\left[\left\|x^{t-\tau} - x^*\right\|^2\right]\\
&+ \sum_{t=\tau}^{T} p_t\varepsilon\gamma\eta L(3p\frac{d}{m}\sqrt{\delta^2+1} + 2\gamma\eta L)\frac{2}{\mu}\,\mathbb{E}[f(x_f^{t-\tau}) - f(x^*)]\\
&+ \sum_{t=\tau}^{T} p_t 2p\gamma\eta\left(\frac{\varepsilon d}{m\sqrt{\delta^2+1}L} + 4p\gamma\eta\frac{d^2}{m^2} + 23p\gamma^3\eta\tau^4\frac{d^3}{m^3}L^2 + p\gamma\eta\tau^2\frac{d^2}{m^2}\left(16\frac{\tau^2d^2L^2p^2\gamma^2}{m^2\beta} + \frac{5}{8}\right)\right)\sigma^2.
\end{aligned}
$$

Similar as in Theorem 15 we take $p_t = p^t$, $p = (1 - \frac{\beta}{2})^{-1}$, it implies $p_\tau \leq 6$ and therefore

$$
\begin{aligned}
&\sum_{t=\tau}^{T} p_t\,\mathbb{E}[\|x^{t+1} - x^*\|^2 + 2\gamma\eta^2(f(x_f^{t+1}) - f(x^*))]\\
&\leq \sum_{t=\tau}^{T} p_t\left(1 - \frac{3\beta}{4} + 6\varepsilon\gamma\eta L\Big(3p\frac{d}{m}\sqrt{\delta^2+1} + 2\gamma\eta L\Big)\right)\mathbb{E}\left\|x^t - x^*\right\|^2\\
&+ \sum_{t=\tau}^{T} p_t\left(2p\gamma\eta^2(1-\frac{p}{\eta}) + 12\frac{\varepsilon\gamma\eta L}{\mu}\Big(3p\frac{d}{m}\sqrt{\delta^2+1} + 2\gamma\eta L\Big)\right)\mathbb{E}[f(x_f^t) - f(x^*)]
\end{aligned}
$$

$$+\sum_{t=\tau}^{T} p_t p^2\gamma^2\eta^2\left(10\frac{d^2}{m^2}(\delta^2+1)-\frac{1}{p}+\tau^2(\delta^2+1)\frac{d^2}{m^2}\left(32\frac{\tau^2d^2L^2p^2\gamma^2}{m^2\beta}+\frac{5}{4}\right)\right)\mathbb{E}\left\|\nabla f(x_g^t)\right\|$$
$$+\sum_{t=0}^{\tau} p_{t+\tau}8p^2\gamma^4\eta^2(\delta^2+1)\frac{d^3}{m^3}\tau^3L^2\left(\frac{2p^2d}{m\beta}+5\right)\sum_{r=t-\tau}^{t-1}\mathbb{E}\left\|\nabla f(x_g^r)\right\|$$
$$+\sum_{t=0}^{\tau} p_{t+\tau}\varepsilon\gamma\eta L(3p\frac{d}{m}\sqrt{\delta^2+1}+2\gamma\eta L)\mathbb{E}\left[\left\|x^t-x^*\right\|^2\right]$$
$$+\sum_{t=0}^{\tau} p_{t+\tau}\varepsilon\gamma\eta L(3p\frac{d}{m}\sqrt{\delta^2+1}+2\gamma\eta L)\frac{2}{\mu}\mathbb{E}[f(x_f^t)-f(x^*)]$$
$$+\sum_{t=\tau}^{T} p_t 2p\gamma\eta\left(\frac{\varepsilon d}{m\sqrt{\delta^2+1}L}+4p\gamma\eta\frac{d^2}{m^2}+23p\gamma^3\eta\tau^4\frac{d^3}{m^3}L^2+p\gamma\eta\tau^2\frac{d^2}{m^2}\left(16\frac{\tau^2d^2L^2p^2\gamma^2}{m^2\beta}+\frac{5}{8}\right)\right)\sigma^2.$$

Taking

$$\gamma\leq\frac{\mu^{\frac{1}{3}}m^{\frac{1}{2}}}{2\tau L^{\frac{4}{3}}d^{\frac{1}{2}}},\quad p\leq\frac{m^2}{13d^2(\delta^2+1)\tau^2},$$
$$\varepsilon\leq\min\left\{\frac{m^{\frac{7}{4}}}{6d^{\frac{7}{4}}\tau^{\frac{5}{4}}L(\delta^2+1)};\frac{m^{\frac{5}{4}}}{\sqrt{2}\tau^{\frac{3}{4}}\mu^{\frac{1}{3}}L^{\frac{2}{3}}d^{\frac{5}{4}}};\frac{m^{\frac{15}{4}}}{6d^{\frac{15}{4}}\tau^{\frac{13}{4}}(\delta^2+1)^2}\right\},$$

we get

$$10\frac{d^2}{m^2}(\delta^2+1)-\frac{1}{p}+\tau^2(\delta^2+1)\frac{d^2}{m^2}\left(32\frac{\tau^2d^2L^2p^2\gamma^2}{m^2\beta}+\frac{5}{4}\right)\leq 0,$$
$$6\varepsilon\gamma\eta L\left(3p\frac{d}{m}\sqrt{\delta^2+1}+2\gamma\eta L\right)\leq\frac{\beta}{4},$$
$$12\frac{\varepsilon\gamma\eta L}{\mu}(3p\frac{d}{m}\sqrt{\delta^2+1}+2\gamma\eta L)\leq 2p\gamma\eta^2\frac{p}{2\eta},$$

and therefore with $\beta=\frac{p}{\eta}$

$$\sum_{t=\tau}^{T} p_t\,\mathbb{E}[\|x^{t+1}-x^*\|^2+2\gamma\eta^2(f(x_f^{t+1})-f(x^*))]\leq\sum_{t=\tau}^{T} p_t\left(1-\frac{\beta}{2}\right)\mathbb{E}[\|x^t-x^*\|^2+2\gamma\eta^2(f(x_f^t)-f(x^*))]$$
$$+\sum_{t=0}^{\tau} p_{t+\tau}8p^2\gamma^4\eta^2(\delta^2+1)\frac{d^3}{m^3}\tau^3L^2\left(\frac{2p^2d}{m\beta}+5\right)\sum_{r=t-\tau}^{t-1}\mathbb{E}\left\|\nabla f(x_g^r)\right\|$$
$$+\sum_{t=0}^{\tau} p_{t+\tau}\varepsilon\gamma\eta L(3p\frac{d}{m}\sqrt{\delta^2+1}+2\gamma\eta L)\mathbb{E}\left[\left\|x^t-x^*\right\|^2\right]$$
$$+\sum_{t=0}^{\tau} p_{t+\tau}\varepsilon\gamma\eta L(3p\frac{d}{m}\sqrt{\delta^2+1}+2\gamma\eta L)\frac{2}{\mu}\mathbb{E}[f(x_f^t)-f(x^*)]$$
$$+\sum_{t=\tau}^{T} p_t 2p\gamma\eta\left(\frac{\varepsilon d}{m\sqrt{\delta^2+1}L}+4p\gamma\eta\frac{d^2}{m^2}+23p\gamma^3\eta\tau^4\frac{d^3}{m^3}L^2+p\gamma\eta\tau\frac{d^2}{m^2}\left(16\frac{\tau^2d^2L^2p^2\gamma^2}{m^2\beta}+\frac{5}{8}\right)\right)\sigma^2.$$

Assume the following notation

$$\Delta_\tau:=\sum_{t=0}^{\tau} p_{t+\tau}8p^2\gamma^4\eta^2(\delta^2+1)\frac{d^3}{m^3}\tau^3L^2\left(\frac{2p^2d}{m\beta}+5\right)\sum_{r=t-\tau}^{t-1}\mathbb{E}\left\|\nabla f(x_g^r)\right\|$$
$$+\sum_{t=0}^{\tau} p_{t+\tau}\varepsilon\gamma\eta L(3p\frac{d}{m}\sqrt{\delta^2+1}+2\gamma\eta L)\mathbb{E}\left[\left\|x^t-x^*\right\|^2\right]$$

$$+\sum_{t=0}^{\tau} p_{t+\tau}\varepsilon\gamma\eta L(3p\frac{d}{m}\sqrt{\delta^2+1}+2\gamma\eta L)\frac{2}{\mu}\,\mathbb{E}[f(x_f^t)-f(x^*)]$$
$$\leq \frac{\sqrt{\gamma}}{\tau^{\frac{4}{3}}\mu^{\frac{1}{3}}}\sum_{t=0}^{\tau}\left(\mathbb{E}\left\|\nabla f(x_g^t)\right\|+\mathbb{E}\left\|x^t-x^*\right\|^2+\mathbb{E}[f(x_f^t)-f(x^*)]\right)$$

Now we substitute p_t, this lead us to

$$\sum_{t=\tau}^{T}\Big(1-\frac{\beta}{2}\Big)^{-t}\mathbb{E}[\|x^{t+1}-x^*\|^2+2\gamma\eta^2(f(x_f^{t+1})-f(x^*))]\leq\sum_{t=\tau}^{T}\Big(1-\frac{\beta}{2}\Big)^{-t+1}\mathbb{E}[\|x^t-x^*\|^2+2\gamma\eta^2(f(x_f^t)-f(x^*))]$$
$$+\Delta_\tau+\sum_{t=\tau}^{T}\Big(1-\frac{\beta}{2}\Big)^{-t}2p\gamma\eta\left(\frac{\varepsilon d}{m\sqrt{\delta^2+1}L}+4p\gamma\eta\frac{d^2}{m^2}+23p\gamma^3\eta\tau^4\frac{d^3}{m^3}L^2+p\gamma\eta\tau\frac{d^2}{m^2}\left(16\frac{\tau^2d^2L^2p^2\gamma^2}{m^2\beta}+\frac{5}{8}\right)\right)\sigma^2.$$

This implies

$$\Big(1-\frac{\beta}{2}\Big)^{-T}\mathbb{E}[\|x^{T+1}-x^*\|^2+2\gamma\eta^2(f(x_f^{T+1})-f(x^*))]\leq\Big(1-\frac{\beta}{2}\Big)^{\tau}\mathbb{E}[\|x^\tau-x^*\|^2+2\gamma\eta^2(f(x_f^\tau)-f(x^*))]+\Delta_\tau$$
$$+\sum_{t=\tau}^{T}\Big(1-\frac{\beta}{2}\Big)^{-t}2p\gamma\eta\left(\frac{\varepsilon d}{m\sqrt{\delta^2+1}L}+4p\gamma\eta\frac{d^2}{m^2}+23p\gamma^3\eta\tau^4\frac{d^3}{m^3}L^2+p\gamma\eta\tau\frac{d^2}{m^2}\left(16\frac{\tau^2d^2L^2p^2\gamma^2}{m^2\beta}+\frac{5}{8}\right)\right)\sigma^2.$$

Rearranging this inequality and taking $\varepsilon\leq\frac{\sqrt{\gamma}m}{\sqrt{\mu}d}$ we obtain

$$\mathbb{E}[\|x^{T+1}-x^*\|^2+2\gamma\eta^2(f(x_f^{T+1})-f(x^*))]\leq\Big(1-\frac{\beta}{2}\Big)^{T-\tau}\mathbb{E}[\|x^\tau-x^*\|^2+2\gamma\eta^2(f(x_f^\tau)-f(x^*))]$$
$$+\Big(1-\frac{\beta}{2}\Big)^{T}\Delta_\tau+6\sqrt{\frac{\gamma}{\mu}}\sigma^2.$$

This finishes the proof. □

H ADDITIONAL EXPERIMENTS

This section provides description of the experiment setup, presents and analyses results of logistic regression experiments on LIBSVM datasets, studies dependence of history size over convergence. Moreover, experiments with neural networks optimization for data-parallelism and model-parallelism are presented and discussed.

H.1 TECHNICAL DETAILS

Our implementation of compression operators and algorithms is written in Python 3.10, with the use of PyTorch optimization library. We implement a simulation of distributed optimization system on a single machine, which is equivalent in terms of convergence analysis. Our server is AMD Ryzen Threadripper 2950X 16-Core Processor @ 2.2 GHz CPU and x2 NVIDIA GeForce GTX 1080 Ti GPU. We use Weights&Biases Biewald [2020] for experiments tracking and hyperparameters tuning.

H.2 LOGISTIC REGRESSION EXPERIMENTS

We conduct experiments on classification with logistic regression on four datasets: `Mushrooms`, `A9A`, `W8A`, `MNIST`. We apply the following optimization algorithms: proposed `MQSGD` and its accelerated version `AMQSGD`, and also use Markovian compressors with popular `DIANA` Mishchenko et al. [2024] algorithm. In all of our experiments, we do not utilize the steps of the optimizer, but rather the information that is transmitted by each worker at the current timestamp t. This implies that there are n workers, with each worker sending m coordinates at each iteration of the optimization step. Consequently, the x-axis displays numbers of the form $mn\cdot 1, mn\cdot 2,\ldots,mn\cdot t,\ldots,mn\cdot T$. This allows us to understand the performance of compressors with varying values of m and n.

We use convex logistic regression loss with a regularization term $\lambda=0.05$. Each dataset is split horizontally (by rows) equally between $N=10$ clients. The feature dimension is denoted as d in the figures, varying from hundreds to almost a

thousand between datasets. The underlying sparsification compressors in Rand-10% for all logistic regression experiments. Learning rate initial value and decay rate are fine-tuned for each problem and compressor. Markovian-specific parameters such as history size K, forgetting rate b are taken from theory or a reasonable default.

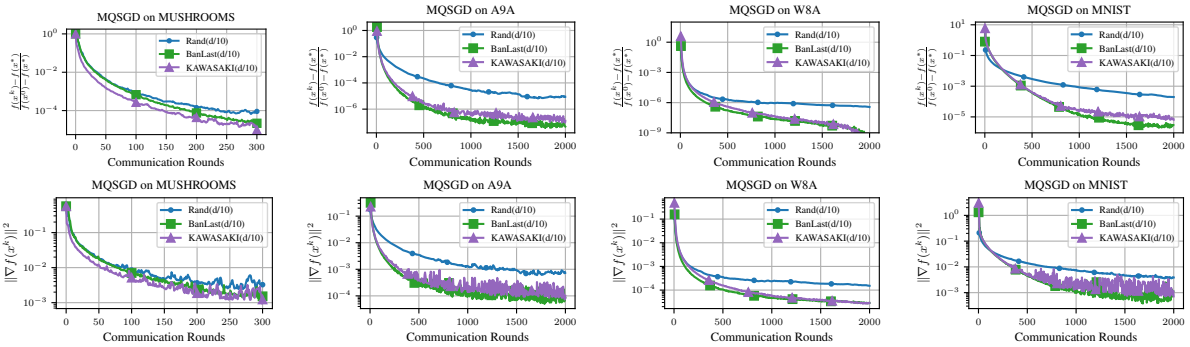

Figure 6: `MQSGD` LIBSVM logistic regression experiments.

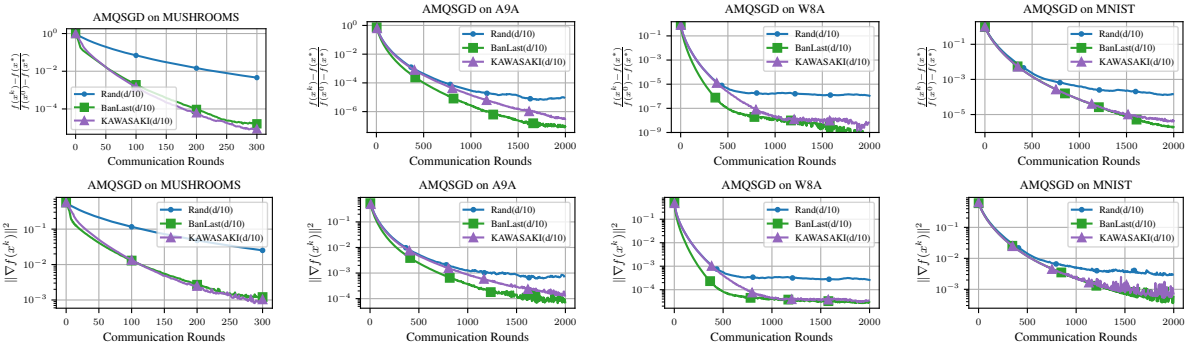

Figure 7: `AMQSGD` LIBSVM logistic regression experiments.

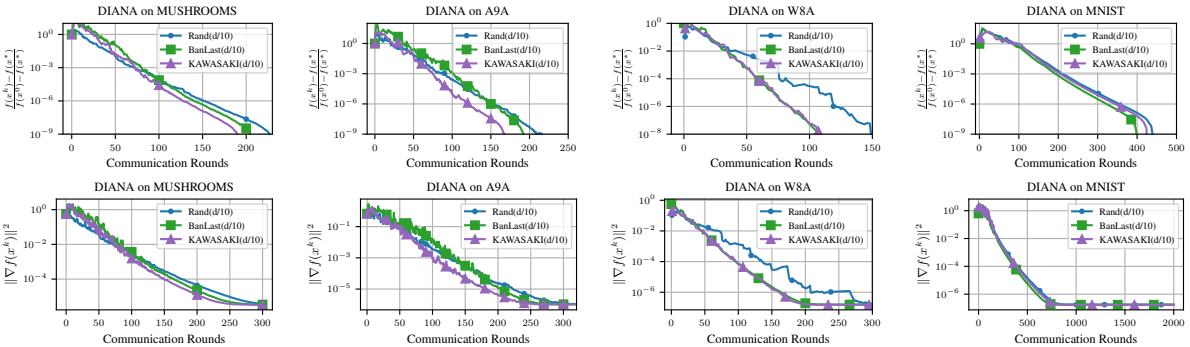

Figure 8: `DIANA` LIBSVM logistic regression experiments.

Figures 6, 7 and 8 present relative distance to the optimum and gradient norm for the best runs on `MQSGD`, `AMQSGD` and `DIANA`, respectively. We observe that Markovian compressors consistently outperform the Rand-10% baseline in all scenarios, as the diverging trend can be seen. Only in some experiments with DIANA (MNIST) the advantage is negligible although present. We also observe that simpler and computational-effective `BanLast` compressor is often enough to achieve substantial convergence improvement. Notably, fine-tuned hyperparameters are similar across datasets and algorithms: for example, `BanLast` tends to perform best with largest possible values of history size K, and `KAWASAKI` forgetting rate b is large. Notice that `BanLast` compressor with largest K turns into round-robin compressor with (almost) no stochasticity in coordinates choice.

H.3 DEPENDENCE ON SIZE HISTORY

As a part of hyperparameter tuning, we additionally analyze how history size K affects the convergence of Markovian compression-based methods. Figure 9 presents dependence of distance to optimum metric on history size for logistic regression experiments. We observe that `BanLast` performs better around larger values of $K = 8$ or $K = 9$. In such case for Rand10% used along with `BanLast`(9), the compression procedure resembles a permutation: for each 10 iterations, no indices are repeated, and the transmission cycle repeats after that. `KAWASAKI` history size seems to have periodical spikes and drops, achieving minimum at around $K = 25$. However, statistics for `DIANA` differ drastically, indicating that history size should be adjusted for each problem independently.

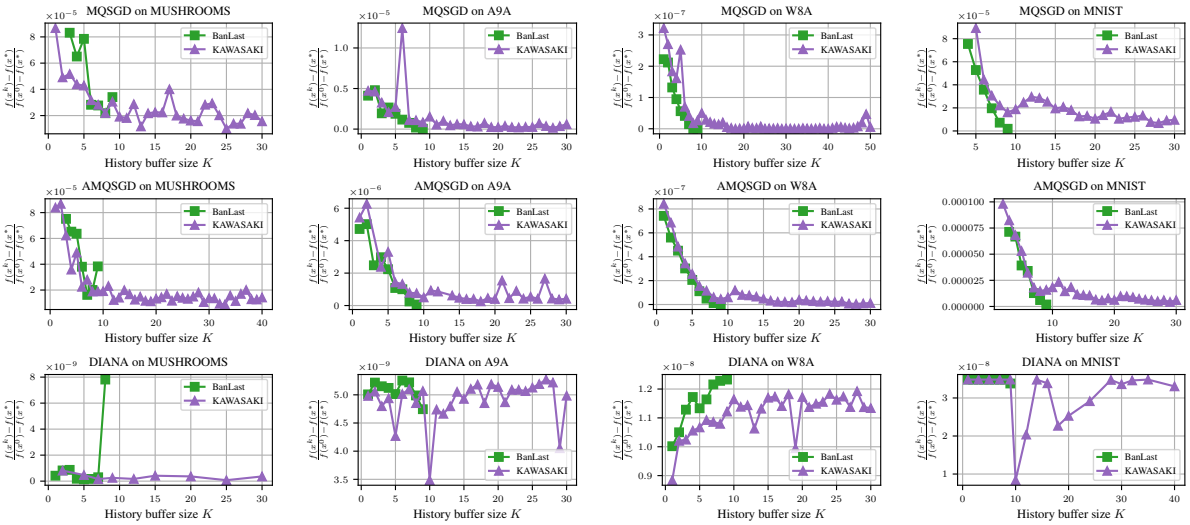

Figure 9: Convergence of Markovian-based algorithms on history size K.

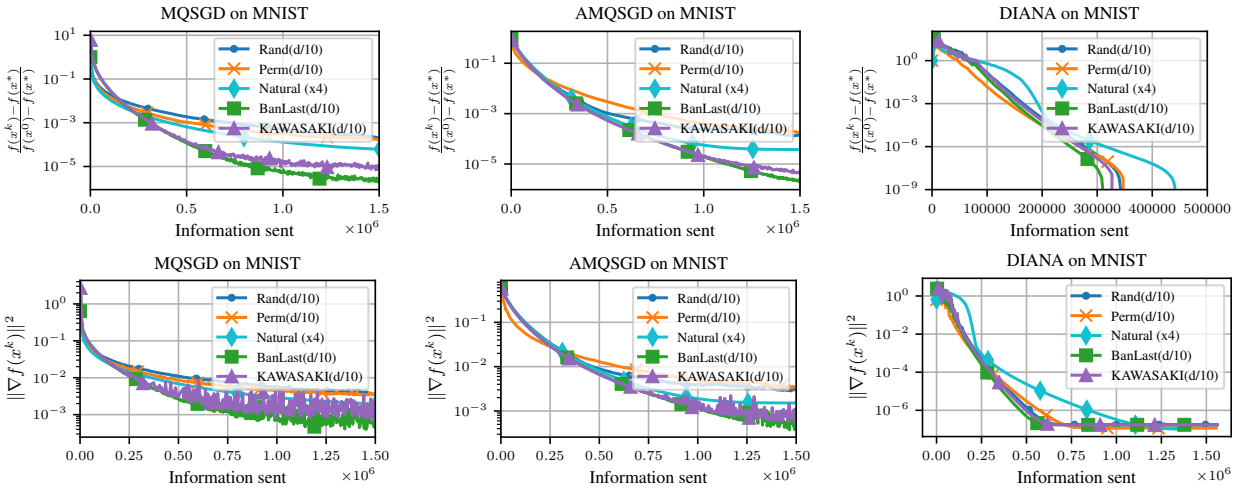

Figure 10: Comparison with PermK compressor and Natural compression.

H.4 COMPARISON WITH PERMUTATION & NATURAL COMPRESSION

In this section, we provide empirical comparison of the proposed compressors with other complex compression schemes. Markovian compressors proposed in the paper compress vector coordinates dependently over optimization epochs. A similar idea of distributed compression is proposed in PermK Szlendak et al. [2022], where coordinates are arranged between workers at each iteration. Another compressor in the consideration is Natural compression Horvath et al. [2022], an unbiased randomized compressor. We perform logistic regression with L2 regularization on MNIST dataset for MQSGD, AMQSGD

and DIANA algorithms on $N = 5$ clients. PermK compression factor is 10, Natural compression factor is 4. Results of comparison of these compressors on MNIST dataset are presented in Figure 10. Best run is shown after fine-tuning learning rate, decay and compression factor. The results justify that Markovian compressors tend to converge faster than the competitors, allowing larger learning rates.

H.5 COMBINATION WITH DIFFERENT COMPRESSORS

Although markovian compressors are initially targeted to work with sparsification-based compressors, refining coordinates selection probabilities, they are fully compatible with other compressors afterwards. To illustrate this, and to conduct additional comparison with PermK compressor, we setup experiments combined with Natural Compression Horvath et al. [2022]. Precisely, we compare RandK+Natural, PermK+Natural, BanLast+Natural and `KAWASAKI`+Natural compressors on logistic regression on MNIST dataset. Figure 11 shows the results of the combination of the mentioned sparsification compressors with natural compression.

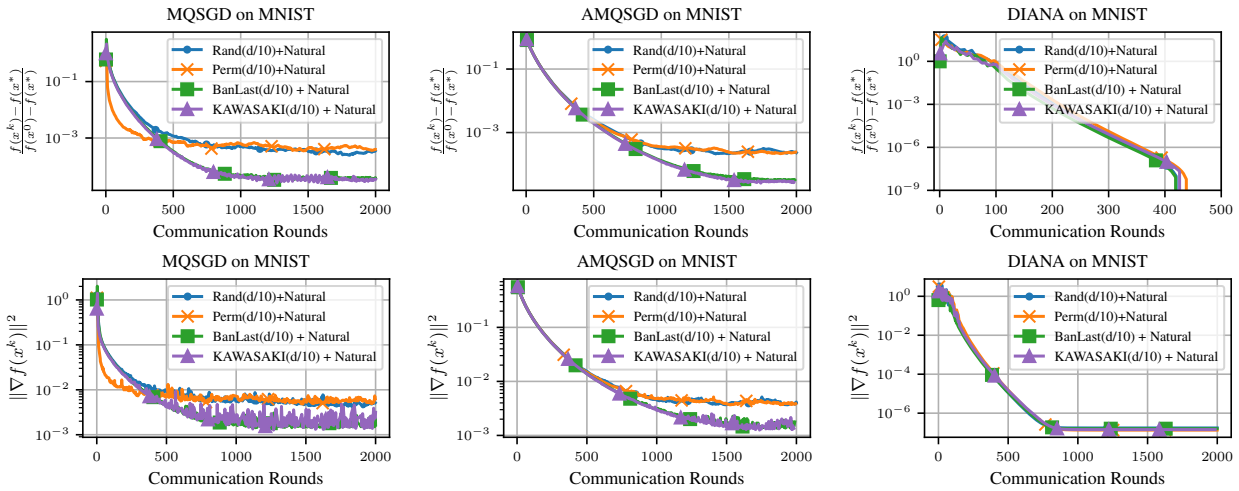

Figure 11: Experiments with Natural compression, MNIST logistic regression experiments.

H.6 NEURAL NETWORKS EXPERIMENTS: DATA PARALLELISM CASE

To adopt Markovian compression to a more complex task, we perform image classification on CIFAR-10 Krizhevsky et al. [2009] with Resnet-18 He et al. [2016] convolutional neural network. We split the training set of size $50,000$ equally between $N = 5$ clients. We use SGD optimizer with momentum 0.9 and weight decay $5 \cdot 10^{-4}$. Hyperparameters such as batch size and learning rate are fine-tuned. Markovian compresors hyperparameters, such as history size K, forgetting rate b, normalization function are initialized with theoretical optimum or a default. Experiments are conducted with several sparsification compressors, such as Rand-5%, Rand-7%, and Rand-10%, with number of epochs adjusted for each case.

Table 2: Best test accuracy % of training ResNet-18 on CIFAR-10 with different compressors

	Rand-K%	`Banlast`	`KAWASAKI`
Rand-5%	88.03	88.1	**89.27**
Rand-7%	89.31	89.38	**90.28**
Rand-10%	**91.46**	**91.72**	**91.78**

Figures 12, 13 and 14 present train loss, gradient norm and test accuracy for each baseline method and Markovian compressors for Rand-5%, Rand-7% and Rand-10% scenarios, respectively. Summary on best test accuracy is presented in Table 2, and extended numerical results for Rand-5% compressor were presented in main experiments Table 1. We observe that in such complex, batched optimization problem only `KAWASAKI` obtains a substantial convergence improvement, as opposed to simpler logistic regression. . In terms of achieved test set accuracy, methods differ significantly only on higher compression rates like Rand-5%. This may imply that Markovian compression tolerates stronger compression, which is

useful in practice. To summarize, Markovian compressors can be successfully applied in neural networks training, with KAWASAKI compressor significantly improving convergence.

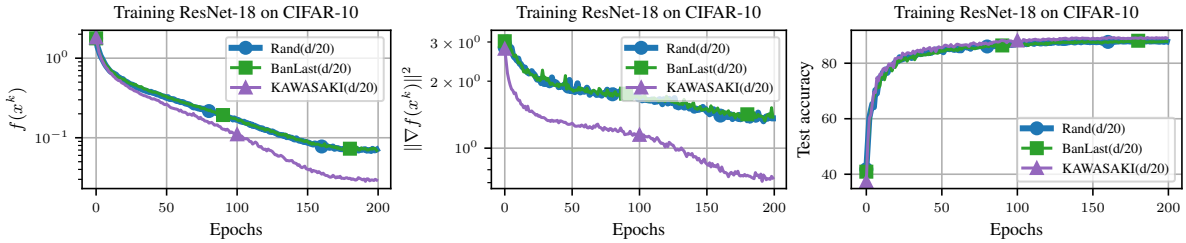

Figure 12: Resnet-18 on CIFAR-10 training results for Rand-5% sparsification.

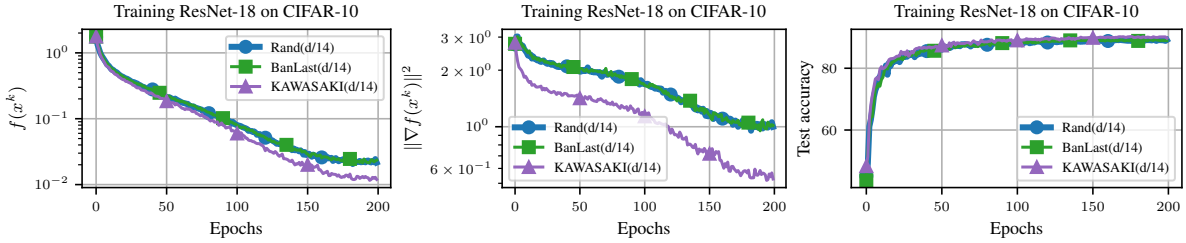

Figure 13: Resnet-18 on CIFAR-10 training results for Rand-7% sparsification.

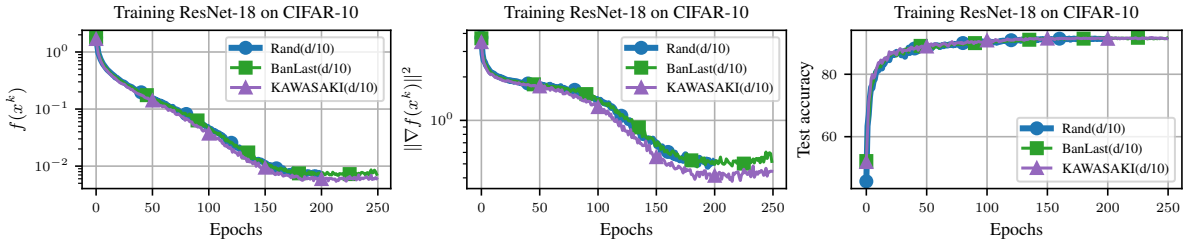

Figure 14: Resnet-18 on CIFAR-10 training results for Rand-10% sparsification.

Finally, we also conduct the comparison with Permutation and Natural compression, both independently and in combination. Figure 15 shows learning curves for training with $N = 20$ clients. KAWASAKI compressor appears to have best convergence in both independently and in combination with Natural compression againt Permutation compressor.

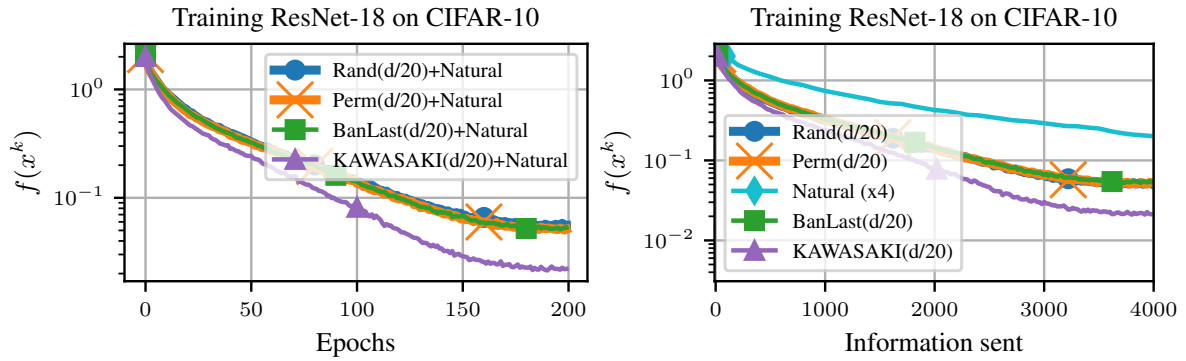

Figure 15: Comparison with other compressors on Resnet-18 training on CIFAR-10 dataset for Rand-5% sparsification on $N = 20$ clients. Natural compression factor is 4. Left figure is sequential combination with Natural compression. Right figure is comparison against PermK and Natural compressors independently, with information sent on x-axis.

H.7 NEURAL NETWORKS EXPERIMENTS: MODEL PARALLELISM CASE

As opposed to data-parallel setting, model parallelism is paradigm which splits the model to a pipeline of layers between workers. This distributed scenario is especially relevant for large language models. As communication is a typical bottleneck in such systems Diskin et al. [2021], various compression techniques are applied to layer activations and their respective gradients that are transferred between adjacent pipeline workers. Such techniques include quantization and sparsification Dettmers et al. [2022], Bian et al. [2023], as well as low-rank compression Song et al. [2023] techniques.

We perform training of Resnet-18 He et al. [2016] convolutional neural network on CIFAR-10 dataset Krizhevsky et al. [2009]. We split the ResNet onto 4 workers by resnet blocks, simulated on a single device with compression of activations and their respective gradients in the places of communication. We apply Markovian compressors only to gradients in model-parallel setup, using same RandK compression for both activations and gradients independently for each compression block.

Table 3: Best test accuracy (%) for model parallelism experiments with Resnet-18 classification of CIFAR-10

Compressor	**Compression ON**	**Compression OFF**
No compression	92.8	92.8
Rand10%	84.6	86.1
BanLastK+Rand10%	85.2	86.4
KAWASAKI(simplex projection)+Rand10%	84.5	85.0
KAWASAKI(normalize)+Rand10%	85.2	**86.8**
KAWASAKI(softmax)+Rand10%	85.3	**87.3**

Table 3 presents best test set accuracy achieved for training with different compressors. While compression indeed decreases accuracy for Rand-10%, application of Markov compressors favors the final test accuracy on a whole one percent. Note that compression is not applied during inference, only on training phase. This case illustrates potential of Markov compressors beyond data-parallelism setup considered in theory.

H.8 FINE-TUNING DEBERTAV3-BASE FOR GLUE BENCHMARK

In this series of experiments, we examine a distributed approach to fine-tuning language models using LoRA [Hu et al., 2022]. This method is based on freezing the model weights that are pre-trained on a large dataset, and add a low rank adapter with matrices $A \in \mathbb{R}^{n\times r}$ and $B \in \mathbb{R}^{r\times m}$ to some selected layers $W_{\text{old}} \in \mathbb{R}^{n\times m}$ of this model, such that $W_{\text{new}} = W_{\text{old}} + A \cdot B$. Since in practice the parameter r is chosen to be much smaller than n and m, the new model has much fewer trainable parameters and can be efficiently trained on downsteram tasks.

In our experiments, we apply LoRA adapters with fixed rank $r = 8$ to the attention layers of the DeBERTaV3-base model [He et al., 2021]. The downsteram task is the classical GLUE benchmark for natural language understanding [Wang et al., 2019]. We consider only random sparsification compressors (Definition 4) with 25% compression rate, due to the large computational cost of this experiment. Figure 16 shows learning curves for training with $N = 10$ clients. Our Markovian compressors appears to have best convergence against independent Rand$m$ compressor.

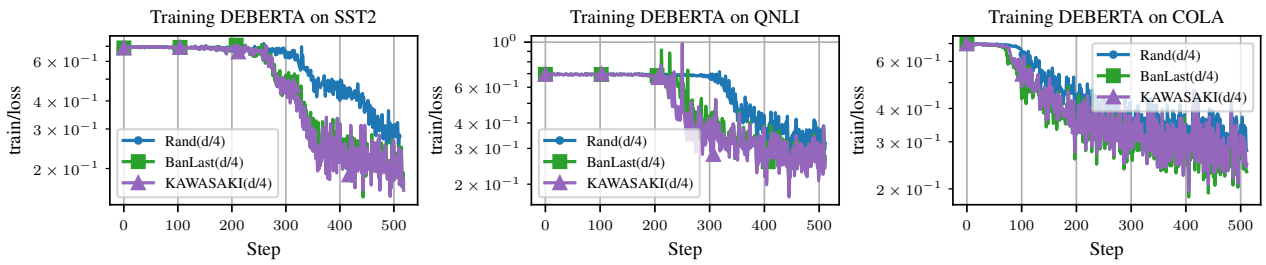

Figure 16: Comparison with other compressors on fine-tuning task on GLUE benchmark on $N = 10$ clients. We performed experiments on SST2, QNLI and COLA tasks.