Artemis: tight convergence guarantees for bidirectional compression in heterogeneous settings for federated learning

Constantin Philippenko, Aymeric Dieuleveut

Artemis: tight convergence guarantees for bidirectional compression in heterogeneous settings for federated learning.

Constantin Philippenko[a,*], Aymeric Dieuleveut[a]

[a] *CMAP, École polytechnique, Institut Polytechnique de Paris, Rte de Saclay, 91120, Palaiseau, France*

Abstract

We introduce a framework – `Artemis` – to tackle the problem of learning in a distributed or federated setting with communication constraints. Several workers (randomly sampled) perform the optimization process using a central server to aggregate their computations. To alleviate the communication cost, `Artemis` allows to compress the information sent in *both directions* (from the workers to the server and conversely) combined with a memory mechanism. It improves on existing algorithms that only consider unidirectional compression (to the server), or use very strong assumptions on the compression operator. We provide fast rates of convergence (linear up to a threshold) under weak assumptions on the stochastic gradients (noise's variance bounded only *at optimal point*) in non-i.i.d. setting, highlight the impact of memory for unidirectional and bidirectional compression, and analyze Polyak-Ruppert averaging. We use convergence in distribution to obtain a *lower bound* of the asymptotic variance that highlights practical limits of compression.

Keywords: Large-scale optimization, Federated learning, Compression, Clients heterogeneity

1. Introduction

In modern large-scale machine learning applications, optimization has to be processed in a distributed fashion, using a potentially large number N in $\mathbb{N}$ of clients. In the data-parallel framework, each client only accesses a fraction of the data: new challenges have arisen, especially when communication constraints between the workers are present.

In this paper, we focus on first-order methods, especially stochastic gradient descent [5, 35] in a centralized framework: a central machine aggregates the computation of the N workers in a synchronized way. This applies to both the *distributed* [e.g. 21] and the *federated learning* [introduced in 19, 26] settings.

Formally, we consider a number of features $d \in \mathbb{N}^*$, and a convex cost function $F : \mathbb{R}^d \to \mathbb{R}$. We want to solve the following convex optimization problem:

$$\min_{w \in \mathbb{R}^d} F(w) \text{ with } F(w) = \frac{1}{N} \sum_{i=1}^{N} F_i(w)\,, \tag{1}$$

where $(F_i)_{i=1}^N$ is a *local* risk function for the model w on the worker i. Especially, in the classical supervised machine learning framework, we fix a loss ℓ and access, on a worker i, n_i observations $(z_k^i)_{1 \le k \le n_i}$ following a distribution D_i. In this framework, F_i can be either the (weighted) local empirical risk, $w \mapsto (n_i^{-1}) \sum_{k=1}^{n_i} \ell(w, z_k^i)$ or the expected risk $w \mapsto \mathbb{E}_{z \sim D_i}[\ell(w, z)]$. At each iteration of the algorithm, each client can get an *unbiased oracle* on the gradient of the function F_i (typically either by choosing uniformly an observation in its dataset or in a *streaming fashion*, getting a new observation at each step).

Our goal is to reduce the amount of information exchanged between workers, to accelerate the learning process, limit the bandwidth usage, and reduce energy consumption. Indeed, the communication cost has been identified as an important bottleneck in the distributed settings [e.g. 40]. In their overview of the federated learning framework, Kairouz et al. [16] also underline in Section 3.5 two possible directions to reduce this cost: (1) compressing communication from workers to the central server (uplink) (2) compressing the downlink communication.

Most of the papers considering the problem of reducing the communication cost [2, 1, 45, 17, 29, 12, 22, 13] only focus on compressing the message sent from the workers to the central node. This direction has the highest potential to reduce the total runtime given that (i) the bandwidth for upload is generally more limited than for download, and that (ii) for some regimes with a large number of workers, the downlink communication, that corresponds to a "one-to-N" communication, may not be the bottleneck compared to the "N-to-one" uplink.

Nevertheless, there are several reasons to also consider downlink compression. First, the difference between upload and download speeds is not significant enough at all to ignore the impact of the downlink direction (see Appendix B for an analysis of bandwidth). If we consider for instance a small number N of workers training a very

*Corresponding author

Email addresses: `constantin.philippenko@gmail.com` (Constantin Philippenko), `aymeric.dieuleveut@gmail.com` (Aymeric Dieuleveut)

heavy model – the size of Deep Learning models generally exceeds hundreds of MB [7, 14] –, the training speed will be limited by the exchange time of the updates, thus using downlink compression is key to accelerating the process. Secondly, in a different framework in which a network of smartphones collaborate to train a large scale model in a federated framework, participants to the training would not be eager to download a hundreds of MB for each update on their phone. Here again, downlink compression appears to be necessary. To encompass all situations, our framework implements compression in either or both directions with possibly different compression levels.

Bidirectional compression (i.e. compressing both uplink and downlink) raises new challenges. In the downlink step, if we compress the *model*, the quantity compressed does *not* tend to zero. Consequently the compression error significantly hinders convergence. To circumvent this problem we compress the *gradient* that may asymptotically approach zero. Prior to this work, bidirectional compression had been considered by Tang et al. [41], Zheng et al. [48], Liu et al. [23], Yu et al. [46]. In particular, Liu et al. [23] developed (concomitantly and independently to our work) an algorithm called `Dore`, which combines error compensation, a memory mechanism, and model compression, and assumes a uniform bound on the gradient variance. In this work, we provide new results on `Dore`-like algorithms, considering a framework *without error-feedback* using tighter assumptions, and quantifying precisely the impact of data heterogeneity on the convergence.

Indeed, we focus on a *heterogeneous* setting: the data distribution depends on each worker (thus non i.i.d.). We *explicitly control the differences between distributions*. In such a setting, the local gradient at the optimal point $\nabla F_i(w_*)$ may not vanish: to get a vanishing compression error, we introduce a "memory" process [29].

Assumptions made on the gradient oracle directly influence the convergence rate of the algorithm: in this paper, we neither assume that the gradients are uniformly bounded [as in 48] nor that their variance is uniformly bounded [Assumption 3, as in 2, 29, 23, 41, 12]: instead we only assume that the variance is bounded by a constant σ_*^2 *at the optimal point* w_*, and provide linear convergence rates up to a threshold *proportional to* σ_*^2 (as in [8, 10] for non distributed optimization). This is a fundamental difference as the variance bound at the optimal point can be orders of magnitude smaller than the uniform bound used in previous work: this is striking when all loss functions have the same critical point, and thus the noise at the optimal point is null! This happens for example in the *interpolation regime*, which has recently gained importance in the machine learning community [4]. As the empirical risk at the optimal point is null or very close to zero, so are all the loss functions with respect to one example. This is often the case in deep learning [e.g., 47] or in large dimension regression [27].

Overall, we make the following contributions:

1. We describe a framework – `Artemis` – **that encompasses 6 algorithms** (with or without up/down compression, with or without memory). We provide and analyze in Theorem 1 a fast rate of convergence – exponential convergence up to a threshold proportional to σ_*^2, the noise at the optimal point –, **obtaining tighter bounds** than in [2, 29].

2. We explicitly tackle heterogeneity using Assumption 4, proving that the limit variance of `Artemis` with memory is independent from the difference between distributions (as for SGD). **This is the first theoretical guarantee for double compression that explicitly quantifies the impact of non i.i.d. data**.

3. In the non-strongly-convex case, we prove the convergence using Polyak-Ruppert averaging in Theorem 2.

4. We prove *convergence in distribution* of the iterates, and subsequently **provide *a lower bounds* on the asymptotic variance.** This sheds light on the limits of (double) compression, which results in an increase of the algorithm's variance, and can thus only accelerate the learning process for *early iterations* and up to a *"moderate" accuracy*. Interestingly, this *"moderate" accuracy* has to be understood with respect to the *reduced* noise σ_*^2.

5. We provide carefully designed experiments to illustrate our theoretical findings. We use both real datasets and synthetic datasets to highlight each of the insights presented in our theorems. See this repository to find the code to reproduce our experiments,

In Table 1, we highlight the main features and assumptions of `Artemis` compared to recent algorithms using compression.

The rest of the paper is organized as follows: in Section 2 we introduce the framework of `Artemis`. In Subsection 2.1 we describe the assumptions, and we review related work in Subsection 2.2. We then give the theoretical results in Section 3, we present experiments in Section 4, and finally, we conclude in Section 5.

2. Problem statement

We consider the problem described in Equation (1). In the convex case, we assume that there exist at least one optimal point which we denote w_*, we also denote $h_*^i = \nabla F_i(w_*)$, for i in $[\![1, N]\!]$. To solve this problem, we rely on a stochastic gradient descent (SGD) algorithm.

A stochastic gradient g_k^i is provided at iteration k in $\mathbb{N}^*$ to the device i in $[\![1, N]\!]$. This function is then evaluated at point w_{k-1}: to alleviate notation, we will use $g_k^i = g_k^i(w_{k-1})$ and $g_{k,*}^i = g_k^i(w_*)$ to denote the stochastic gradient vectors at points w_{k-1} and w_* on device i. In the classical centralized framework (without compression), SGD corresponds to:

Table 1: Comparison of frameworks for main algorithms handling (bidirectional) compression. By "non i.i.d.", we mean that the theoretical framework encompasses *and* explicitly quantifies the impact of data heterogeneity on convergence (Assumption 4), e.g., `Dore` does not assume i.i.d. workers but does not quantify differences between distributions. References: see [2] for `QSGD`, [29] for `Diana`, [13] for [HR20], [23] for `Dore` and [41] for `DoubleSqueeze`.

	`QSGD`	`Diana`	[HR20]	`Dore`	`Double Squeeze`	`Dist EF-SGD`	`Artemis` **(new)**
Data	i.i.d.	non i.i.d.	non i.i.d.	i.i.d.	i.i.d.	i.i.d.	non i.i.d.
Bounded variance	Uniformly	Uniformly	Uniformly	Uniformly	Uniformly	Uniformly	At optimal point
Compression	One-way	One-way	One-way	Two-way	Two-way	Two-way	Two-way
Error-feedback			✓	✓	✓	✓	
Memory		✓		✓			✓
Partial part.			✓				✓

$$w_k = w_{k-1} - \gamma \frac{1}{N} \sum_{i=1}^{N} g_k^i \tag{2}$$

where γ is the learning rate.

However, computing such a sequence would require the nodes to send either the gradient g_k^i or the updated local model to the central server (*uplink* communication), and the central server to broadcast back either the averaged gradient g_k or the updated global model (*downlink* communication). Here, in order to reduce communication cost, we perform a *bidirectional* compression. More precisely, we combine two main tools: (1) an *unbiased compression operator* $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ that reduces the number of bits exchanged, and (2) a *memory* process that reduces the size of the signal to compress, and consequently the error [29, 22]. That is, instead of directly compressing the gradient, we first approximate it by the memory term and, afterwards, we compress the difference. As a consequence, the compressed term tends in expectation to zero, and the error of compression is reduced. Following Tang et al. [41], we always broadcast gradients and never models. To distinguish the two compression operations we denote $\mathcal{C}_{\text{up}}$ and $\mathcal{C}_{\text{dwn}}$ the compression operator for uplink and downlink. At each iteration, we thus have the following steps:

1. First, each active local node sends to the central server a compression of gradient differences: $\widehat{\Delta}_k^i = \mathcal{C}_{\text{up}}(g_k^i - h_{k-1}^i)$, and updates the *memory term* $h_k^i = h_{k-1}^i + \alpha_{\text{up}} \widehat{\Delta}_k^i$ with $\alpha_{\text{up}} \in \mathbb{R}^*$. The server recovers the approximated gradients' values by adding the received term to the memories kept on its side.

2. Then, the central server sends back the compression of the sum of compressed gradients: $\Omega_k = \mathcal{C}_{\text{dwn}}(\frac{1}{N}\sum_{i=1}^N \widehat{\Delta}_k^i + h_{k-1}^i)$. No memory mechanism needs to be used, as the sum of gradients tends to zero in the absence of regularization.

The update is thus given by:

$$\begin{cases} \forall i \in [\![1, N]\!], \quad \widehat{\Delta}_k^i = \mathcal{C}_{\text{up}}\left(g_k^i - h_{k-1}^i\right) \\ \Omega_k = \mathcal{C}_{\text{dwn}}\left(\frac{1}{N}\sum_{i=1}^N (\widehat{\Delta}_k^i + h_{k-1}^i)\right) \\ w_k = w_{k-1} - \gamma \Omega_k \,. \end{cases} \tag{3}$$

Constants $\gamma, \alpha_{\text{up}} \in \mathbb{R}^* \times \mathbb{R}_+$ are learning rates for respectively the iterate sequence and the memory sequence.

Partial participation.. An important setting of FL is the partial participation (PP) of clients at each round: clients only participate in a fraction p of the training steps. This can be addressed theoretically by modelizing it as a compression scheme $\mathcal{C}_{\text{PP}}$, which compresses a vector z as either z/p or 0. As such, our analysis of uplink compression naturally encompasses the PP scenario. In the PP setting, the main difficulty is to keep all clients synchronized when they return to the training process. This requires sharing any updates they missed or the latest iterate, depending on which option is more efficient. This step is commonly referred to as a "catching-up" process. This approach has also been proposed by Sattler et al. [36, see the remark preceding Equation (20) in Section VI.C] or by Tang et al. [41, v2 on arxiv for the distributed case], who use a buffer. We present the pseudo-code of `Artemis` with the catching-up step in Algorithm 1.

As a summary, the `Artemis` framework encompasses, in particular, these four algorithms: the variant with unidirectional compression ($\omega_{\text{dwn}} = 0$) w.o. or with memory ($\alpha_{\text{up}} = 0$ or $\alpha_{\text{up}} \neq 0$) recovers `QSGD` defined by Alistarh et al. [2] and `DIANA` proposed by [29]. The variant using bidirectional compression ($\omega_{\text{dwn}} \neq 0$) w.o memory ($\alpha_{\text{up}} = 0$) is called `Bi-QSGD`. The last and most effective variant combines bidirectional compression *with* memory and is the one we refer to as `Artemis` if no precision is given. It corresponds to a simplified version of `Dore` without error-feedback, but this additional mechanism did not lead to any theoretical improvement in the case of unbiased compressors [Remark 2 in Sec. 4.1., 23].

Remark 1 (Local steps)**.** *An obvious independent direction to reduce communication is to increase the number of steps performed before communication. This is the spirit of Local-SGD [38]. It is an interesting extension to incorporate this into our framework. We do not consider it in order to focus on the compression insights.*

In the following section, we present and discuss assumptions over the function F, the data distribution and the compression operator.

Algorithm 1: `Artemis` - set $\alpha > 0$ to use memory.

Input: Mini-batch size b, learning rates $\alpha, \gamma > 0$, initial model $w_0 \in \mathbb{R}^d$, operators $\mathcal{C}_{\text{up}}$ and $\mathcal{C}_{\text{dwn}}$, M_1 and M_2 the sizes of the full/compressed gradients.

Initialization: Local memory: $\forall i \in [\![1, N]\!]\ h_0^i = 0$ (kept on both central server and device i). Index of last participation: $k_i = 0$.

Output: Model w_K

for $k = 0, 1, 2, \dots, K$ **do**

Randomly sample a set of device S_k

for *each device* $i \in S_k$ **do**

Catching up.

If $k - k_i > \lfloor M_1/M_2 \rfloor$, send the model w_k

Else send $(\widehat{\Omega}_j)_{j=k_i+1}^k$ and update local model: $\forall j \in [\![k_i + 1, k]\!], \quad w_j = w_{j-1} - \gamma \Omega_{j,S_{j-1}}$

Update index of its last participation: $k_i = k$

Local training.

Compute stochastic gradient $g_k^i = g_{k+1}(w_k)$ (with mini-batch)

Set $\Delta_k^i = g_k^i - h_k^i$, compress it $\widehat{\Delta}_k^i = \mathcal{C}_{\text{up}}(\Delta_k^i)$

Update memory term: $h_{k+1}^i = h_k^i + \alpha \widehat{\Delta}_k^i$

Send $\widehat{\Delta}_k^i$ to central server

Compute $\widehat{g}_k = h_k + \frac{1}{pN} \sum_{i \in S_k} \widehat{\Delta}_k^i$

Update central memory: $h_{k+1} = h_k + \alpha \frac{1}{N} \sum_{i \in S_k} \widehat{\Delta}_k^i$

Back compression: $\Omega_{k+1,S_k} = \mathcal{C}_{\text{dwn}}(\widehat{g}_k)$

Broadcast Ω_{k+1} to all workers.

Update model on central server: $w_{k+1} = w_k - \gamma \Omega_{k+1,S_k}$

2.1. Assumptions

We make classical assumptions on $F : \mathbb{R}^d \to \mathbb{R}$.

Assumption 1 (Strong-convexity)**.** *F is μ-strongly-convex, that is for all vectors z, z' in $\mathbb{R}^d$: $F(z') \geq F(w) + (z' - z)^T \nabla F(z) + \frac{\mu}{2} \|z' - z\|_2^2$.*

Note that we do not need each F_i to be strongly convex, but only F. Also remark that we only use this inequality for $z' = w_*$ in the proof of Theorems 1 and 2.

Below, we introduce cocoercivity [see 50, for more details about this hypothesis]. This assumption implies that all $(F_i)_{i \in [\![1,N]\!]}$ are L-smooth.

Assumption 2 (Cocoercivity of stochastic gradients in quadratic mean)**.** *We suppose that for all k in $\mathbb{N}$, stochastic gradients functions $(g_k^i)_{i \in [\![1,N]\!]}$ are L-cocoercive in quadratic mean. That is, for k in $\mathbb{N}$, i in $[\![1, N]\!]$ and for all vectors z, z' in $\mathbb{R}^d$, we have:*

$$\mathbb{E}[\|g_k^i(z) - g_k^i(z')\|^2] \leq L \left\langle \nabla F_i(z) - \nabla F_i(z'), z - z' \right\rangle .$$

E.g., this is true under the much stronger assumption that stochastic gradients functions $(g_k^i)_{i \in [\![1,N]\!]}$ are *almost surely* L-cocoercive, i.e.:

$$\|g_k^i(z) - g_k^i(z')\|^2 \leq L \left\langle g_k^i(z) - g_k^i(z'), z - z' \right\rangle .$$

Next, we present the assumption on the stochastic gradient's noise. Again, we highlight that the noise is only controlled at the optimal point. To carefully control the noises process (gradient oracle, uplink, and downlink compression), we introduce three filtrations $(\mathcal{H}_k, \mathcal{G}_k, \mathcal{F}_k)_{k \geq 0}$, such that w_k is $\mathcal{H}_k$-measurable for any $k \in \mathbb{N}$. Detailed definitions are given in Appendix A.

Assumption 3 (Noise over stochastic gradients computation)**.** *The noise over stochastic gradients at the global optimal point, for a mini-batch of size b, is bounded: there exists a constant $\sigma_* \in \mathbb{R}$, s. t. for all k in $\mathbb{N}$, for all i in $[\![1, N]\!]$, we have a.s.: $\mathbb{E}[\|g_{k,*}^i - \nabla F_i(w_*)\|^2 | \mathcal{H}_{k-1}] \leq \frac{\sigma_*^2}{b}$.*

In fact, Assumption 3 only requires that for any $i \in [\![1, N]\!]$, $\mathbb{E}[\|g_{k,*}^i - \nabla F_i(w_*)\|^2 \mid \mathcal{H}_{k-1}] \leq \frac{\sigma_{*,i}^2}{b}$, and the results then hold for $\sigma_*^2 = \frac{1}{N} \sum_{i=1}^N \sigma_{*,i}^2$. In other words, the bounds do not need to be uniform over workers, only the average truly matters. The constant σ_*^2 is null, for example, if we use deterministic (batch) gradients, *or* in the interpolation regime for i.i.d. observations, as discussed in the Introduction. As we have also incorporated here a mini-batch parameter, this reduces the variance by a factor b.

Unlike `Diana` [29, 22], `Dore` [23], `Dist-EF-SGD` [48] or `Double-Squeeze` [41], we assume that the variance of the noise is bounded *only at optimal point* w_* and not *at any point* w in $\mathbb{R}^d$. It results that if the variance is null ($\sigma_*^2 = 0$) at the optimal point, we obtain a linear convergence while previous results obtain this rate solely if the variance is null *at any point* (i.e. only for deterministic GD). Also remark that Assumptions 2 and 3 both stand for the simplest

Least-Squares Regression (LSR) setting, while the uniform bound on the gradient's variance *does not*. Next, we give the assumption that links the distributions on the different machines.

Assumption 4 (Bounded gradient at w_*)**.** *There exists a constant $B \in \mathbb{R}_+$, s.t.:*

$$\frac{1}{N}\sum_{i=0}^{N} \|\nabla F_i(w_*)\|^2 = B^2 \,.$$

This assumption is used to quantify how different the distributions are on the different clients. In the streaming *i.i.d. setting* – $D_1 = \cdots = D_N$ and $F_1 = \cdots = F_N$ – the assumption is satisfied with $B = 0$. Combining Assumptions 3 and 4 results in an upper bound on the averaged squared norm of stochastic gradients at w_*: for all k in $\mathbb{N}$, we have a.s. $\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[\|g_{k,*}^i\|^2 | \mathcal{H}_{k-1}] \leq \frac{\sigma_*^2}{b} + B^2$.

Finally, compression operators can be classified in two main categories: quantization [as in 31, 2, 37, 49, 44, 34, 12] and random projection [as in 42, 32, 39, 3, 18]. Theoretical guarantees provided in this paper do not rely on a particular kind of compression, as we only consider the following assumption on the compression operators $\mathcal{C}_{\text{up}}$ and $\mathcal{C}_{\text{dwn}}$:

Assumption 5. *There exist constants $\omega_{\text{up}}, \omega_{\text{dwn}} \in \mathbb{R}_+^*$, such that the compression operators $\mathcal{C}_{\text{up}}$ and $\mathcal{C}_{\text{dwn}}$ verify the two following properties for all z in $\mathbb{R}^d$:*

$$\begin{cases} \mathbb{E}[\mathcal{C}_{\text{up/dwn}}(z)] = z \,, \\ \mathbb{E}[\left\|\mathcal{C}_{\text{up/dwn}}(z) - z\right\|^2] \leq \omega_{\text{up/dwn}} \|z\|^2 \,. \end{cases}$$

In other words, the compression operators are unbiased and their variances are relatively bounded. Assumption 5 requires in fact to access a *sequence of i.i.d. compression operators* $\mathcal{C}_{\text{up/dwn},k}$ for $k \in \mathbb{N}$ – but for simplicity, we generally omit the k index. Note that Horváth & Richtárik [13] have shown that using an unbiased operator leads to better performances. Unlike us, Tang et al. [41] assume uniformly bounded compression error, which is a much more restrictive assumption. Also note that $\omega_{\text{up/dwn}}$ can be considered as *parameters* of the algorithm, as the compression levels can be chosen. We now provide additional details on related papers dealing with compression.

2.2. Related work on compression

Quantization is a common method for compression and is used in various algorithms. For instance, Seide et al. [37] are one of the first to propose to quantize each gradient component by either -1 or 1. This approach has been extended in Karimireddy et al. [17]. Alistarh et al. [2] define a new algorithm – `QSGD` – which instead of sending gradients, broadcasts their quantized version, getting robust results with this approach. On top of gradient compression, Wu et al. [45] add an error-compensation mechanism that accumulates quantization errors and corrects the gradient computation at each iteration. `Diana` [introduced in 29] introduces a "memory" term in the place of accumulating errors. Li et al. [22] extend this algorithm and improve its convergence by using an accelerated gradient descent. Reisizadeh et al. [34] combine unidirectional quantization with device sampling, leading to a framework closer to federated learning settings where devices can easily be switched off. In the same perspective, Horváth & Richtárik [13] detail results that also consider PP. Tang et al. [41] are the first to suggest a bidirectional compression scheme for a decentralized network. For both uplink and downlink, the method consists in sending a compression of gradients combined with an error compensation. Later, Yu et al. [46] choose to compress models instead of compressing gradients. This approach is enhanced by Liu et al. [23] who combine model compression with a memory mechanism and an error compensation drawing from [29]. Both Tang et al. [41] and Zheng et al. [48] compress gradients without using a memory mechanism. However, as proved in the following section, memory is key to reducing the asymptotic variance in the heterogeneous case. We now provide theoretical results about the convergence of bidirectional compression.

Table 2: Details on constants C and E defined in Theorem 1. $C = 0$ for $\alpha_{\text{up}} = 0$, see Th. S5 for $\alpha_{\text{up}} \neq 0$.

α_{up}	E
0	$(\omega_{\text{dwn}}+1)\left((\omega_{\text{up}}+1)\frac{\sigma_*^2}{b} + \omega_{\text{up}} B^2\right)$
$\neq 0$	$\frac{\sigma_*^2}{b}\left((2\omega_{\text{up}}+1)(\omega_{\text{dwn}}+1) + 4\alpha_{\text{up}}^2 C(\omega_{\text{up}}+1) - 2\alpha_{\text{up}} C\right)$

3. Theoretical results

In this Section, we present our main theoretical results on the convergence of `Artemis` and its variants. To ensure clarity, the most complete and tightest versions of theorems are given in Appendices, while offering simplified versions here.

The main linear convergence rates are given in Theorem 1, and in Theorem 2 we show that `Artemis` combined with Polyak-Ruppert averaging reaches a sub-linear convergence rate. We denote $\delta_0^2 = \|w_0 - w_*\|^2$.

Theorem 1 (Convergence of `Artemis`)**.** *Under Assumptions 1 to 5, for a step-size γ satisfying the conditions in Table 3, for a learning rate α_{up} and for any k in $\mathbb{N}$, the mean squared distance to w_* decreases at a linear rate up to a constant of the order of E:*

$$\mathbb{E}\left[\|w_k - w_*\|^2\right] \leq (1-\gamma\mu)^k \left(\delta_0^2 + 2C\gamma^2 B^2\right) + \frac{2\gamma E}{\mu N} \,,$$

for constants C and E depending on the variant (independent of k) given in Table 2 or in the appendix. Variants with $\alpha_{\text{up}} \neq 0$ require $\alpha_{\text{up}} \in [1/2(\omega_{\text{up}}+1), \alpha_{\max}]$, the upper bound $\alpha_{\max}$ is given in Theorem S5.

This theorem is derived from Theorems S4 and S5 which are respectively proved in Appendices E.1 and E.2.

We can make the following remarks:

1. **Linear convergence.** The convergence rate given in Theorem 1 can be decomposed into two terms: a bias term, forgotten at linear speed $(1-\gamma\mu)^k$, and a variance residual term which corresponds to the *saturation level* of the algorithm. The rate of convergence $(1-\gamma\mu)$ does not depend on the variant of the algorithm. However, the variance and initial bias do vary.

2. **Bias term.** The initial bias always depends on $\|w_0 - w_*\|^2$, and when using memory (i.e. $\alpha_{\rm up} \neq 0$) it also depends on the difference between distributions (constant B^2).

3. **Variance term and memory.** On the other hand, the variance depends (1) on both σ_*^2/b, and the distributions' difference B^2 without memory (2) only on the gradients' variance *at the optimum* σ_*^2/b with memory. Similar theorems in related literature [23, 2, 29, 46, 41, 48] systematically had a worse bound for the variance term depending on a *uniform bound of the noise variance* or under much stronger conditions on the compression operator. This work and [23] are also the first to give a linear convergence up to a threshold for bidirectional compression.

4. **Impact of memory.** To the best of our knowledge, this is the first work on double compression that explicitly tackles the non i.i.d. case. We prove that memory makes the saturation threshold independent of B^2 for `Artemis`.

5. **Variance term.** The variance term increases with a factor proportional to $\omega_{\rm up}$ for the unidirectional compression, and proportional to $\omega_{\rm up} \times \omega_{\rm dwn}$ for bidirectional. This is the counterpart of compression, each compression resulting in a multiplicative factor on the noise. A similar increase in the variance appears in [29] and [23]. The noise level is attenuated by the number of devices N, to which it is inversely proportional.

6. **Link with classical SGD.** For variant of `Artemis` with $\alpha_{\rm up} = 0$, if $\omega_{\rm up/dwn} = 0$ (i.e. no compression) we recover SGD results: convergence does not depend on B^2, but only on the noise's variance.

Conclusion: Overall, it appears that `Artemis` is able to efficiently accelerate the learning during first iterations, enjoying the same linear rate as SGD with lower communication complexity, but it saturates at a higher level, proportional to σ_*^2 and independent of B^2.

The range of acceptable learning rates is an important feature for first order algorithms. In Table 3, we summarize the upper bound $\gamma_{\max}$ on γ, to guarantee a $(1-\gamma\mu)$ convergence of `Artemis`. These bounds are derived from Theorems S4 and S5, in three main asymptotic regimes: $N \gg \omega^{\rm up}$, $N \approx \omega^{\rm up}$ and $\omega^{\rm up} \gg N$. Using bidirectional compression impacts $\gamma_{\max}$ by a factor $\omega_{\rm dwn}+1$ in comparison to unidirectional compression. For unidirectional compression, if the number of machines is at least of the order of $\omega_{\rm up}$, then $\gamma_{\max}$ nearly corresponds to $\gamma_{\max}$ for vanilla (serial) SGD.

Table 3: Upper bound on $\gamma_{\max}$ to guarantee convergence. For unidirectional compression (resp. no compr.), $\omega_{\rm dwn} = 0$ (resp. $\omega_{\rm up/dwn} = 0$, recovering classical rates for SGD).

Memory	$\alpha_{\rm up} = 0$	$\alpha_{\rm up} \neq 0$
$N \gg \omega_{\rm up}$	$\frac{1}{(\omega_{\rm dwn}+1)L}$	$\frac{1}{2(\omega_{\rm dwn}+1)L}$
$N \approx \omega_{\rm up}$	$\frac{1}{3(\omega_{\rm dwn}+1)L}$	$\frac{1}{5(\omega_{\rm dwn}+1)L}$
$\omega_{\rm up} \gg N$	$\frac{N}{2\omega_{\rm up}(\omega_{\rm dwn}+1)L}$	$\frac{N}{4\omega_{\rm up}(\omega_{\rm dwn}+1)L}$

We now provide a convergence guarantee for the averaged iterate without strong-convexity.

Theorem 2 (Convergence of `Artemis` with Polyak-Ruppert averaging)**.** *Under Assumptions 2 to 5 (convex case) with constants C and E as in Theorem 1 (see Table 2 for precision), after running K in $\mathbb{N}^*$ iterations, for a learning rate $\gamma = \min\left(\sqrt{\frac{N\delta_0^2}{2EK}}; \gamma_{\max}\right)$, with $\gamma_{\max}$ as in Table 3, we have a sublinear convergence rate for the Polyak-Ruppert averaged iterate $\overline{w}_{K-1} = \frac{1}{K}\sum_{k=0}^{K-1} w_k$:*

$$\mathbb{E}\left[F(\overline{w}_{K-1}) - F(w_*)\right] \leq 2\max\left(\sqrt{\frac{2\delta_0^2 E}{NK}}; \frac{\delta_0^2}{\gamma_{\max}K}\right) + \frac{2\gamma_{\max}CB^2}{K}.$$

This theorem is proved in Appendix E.3. Several comments can be made on this theorem:

1. **Importance of averaging** This is the first theorem given for averaging for double compression. In the context of convex optimization, averaging has been shown to be optimal [33].

2. **Speed of convergence, if $\sigma_* = 0$, $B \neq 0$, $K \to \infty$.** For $\alpha_{\rm up} \neq 0$, $E = 0$, while for $\alpha_{\rm up} = 0$, $E \propto B^2$. Memory thus accelerates the convergence from a rate $O(K^{-1/2})$ to $O(K^{-1})$.

3. **Speed of convergence, general case.** More generally, we always get a $K^{-1/2}$ sublinear speed of convergence, and a faster rate K^{-1} when using memory and if $E \leq \delta_0^2 N/(2K\gamma_{\max}^2)$ – i.e. in the context of a low noise σ_*^2, as $E \propto \sigma_*^2$. Again, it appears that bi-compression is mostly useful in low-σ_*^2 regimes or during the first iterations: intuitively, for a fixed communication budget, while bi-compression allows to perform $\min\{\omega_{\rm up}, \omega_{\rm dwn}\}$-times more iterations, this is no longer beneficial if the convergence rate is dominated by $\sqrt{2\delta_0^2 E/NK}$, as E increases by a factor $\omega_{\rm up} \times \omega_{\rm dwn}$.

4. **Memoryless case, impact of minibatch.** In the variant of `Artemis` *without memory*, the asymptotic convergence rate is $\sqrt{2\delta_0^2 E/NK}$ with the constant $E \propto$

$\sigma_*^2/b + B^2$: interestingly, it appears that in the case of non i.i.d. data ($B^2 > 0$), the *convergence rate saturates when the size of the mini-batch increases*: large mini-batches *do not help*. On the contrary, with memory, the variance is, as classically, reduced by a factor proportional to the size of the batch, without saturation.

The increase in the variance (in Item 3) is not an artifact of the proof: indeed we provide a corresponding (algorithm-specific) lower bound based on proving the existence of a limit distribution for the iterates of `Artemis`, and analyzing its variance, see Theorem 3 in next Subsection.

3.1. Convergence in distribution and lower bound

The increase in the variance (in Item 3) is not an artifact of the proof: we prove the existence of a limit distribution for the iterates of `Artemis`, and analyze its variance. More precisely, we show a linear rate of convergence for the distribution Θ_k of w_k (launched from w_0), w.r.t. the Wasserstein distance $\mathcal{W}_2$ [43]: this gives us a lower bound on the asymptotic variance. Here, we further assume that the compression operator is *linear* (e.g., sparsification, sketching, rand-h, PP).

Theorem 3 (Convergence in distribution and lower bound on the variance)**.** *Under Assumptions 1 to 5, for $\gamma, \alpha_{\rm up}, E$ given in Theorem 1 and Table 3:*

1. *There exists a limit distribution $\pi_{\gamma,v}$ depending on the variant v of the algorithm, s.t. for any $k \geq 1$, we have $\mathcal{W}_2(\Theta_k, \pi_{\gamma,v}) \leq (1-\gamma\mu)^k C_0$, with C_0 a constant.*
2. *When k goes to infinity, the second order moment $\mathbb{E}[\|w_k - w_*\|^2]$ converges to $\mathbb{E}_{w\sim\pi_{\gamma,v}}[\|w - w_*\|^2]$, which is* lower bounded *by $\Omega(\gamma E/\mu N)$ as in Theorem 1 as $\gamma \to 0$, with E depending on the variant.*

Interpretation. The second point (2.) means that the upper bound on the saturation level provided in Theorem 1 *is tight* w.r.t. $\sigma_*^2, \omega_{\rm up}, \omega_{\rm dwn}, B^2, N$ and γ. Especially, it proves that there is indeed a quadratic increase in the variance w.r.t. $\omega_{\rm up}$ and $\omega_{\rm dwn}$ when using bidirectional compression (which is itself rather intuitive). Altogether, these three theorems prove that bidirectional compression can become strictly worse than usual stochastic gradient descent in high precision regimes, a fact of major importance in practice and barely (if ever) even mentioned in previous literature. To the best of our knowledge, only [25] are giving a lower bound on the asymptotic variance for algorithms using compression. Their result is more general, i.e., valid for any algorithm using unidirectional compression, but weaker (worst case on the oracle does not highlight the importance of noise at the optimal point and is incompatible with linear rates).

Proof and assumptions. This theorem also naturally requires, for the second point, Assumptions 3 to 5 to be "tight": that is, e.g., $\mathrm{Var}(g_{k,*}^i) \geq \Omega(\sigma_*^2/b)$; more details and the proof are given in Appendix E.4. Extension to other types of compression reveals to be surprisingly non-simple, and is thus out of the scope of this paper and a promising direction.

4. Experiments

In this Section, we illustrate our theoretical guarantees on both synthetic and real datasets, confirm the theoretical findings in Theorems 1 to 3, and underline the impact of the memory. Therefore, we focus on five of the algorithms covered by our framework: `Artemis` with bidirectional compression (simply denoted `Artemis`), `QSGD`, `Diana`, `Bi-QSGD`, and usual `SGD` without any compression. In Appendix C, we provide additional details and more illustrative experiments. In particular, we compare `Artemis` with other existing benchmarks (Figure S17): `Double-Squeeze`, `Dore`, `FedSGD` and `FedPAQ` [see 34]. We also perform experiments with *optimized* learning rates (Figure S16).

In all experiments, we display the logarithm excess loss $\log_{10}(F(w_{k-1}) - F(w_*))$ w.r.t. the number of iterations k or the number of communicated bits. Curves are averaged over 5 runs, and we plot error bars on all figures. These errors bars correspond to $\pm$ the standard deviation of the logarithm excess loss over the five runs. We use a quantization scheme with $s = 2^0$.

Definition 1 (s-quantization operator)**.** *Given $z \in \mathbb{R}^d$, the s-quantization operator $\mathcal{C}_s$ is defined by $\mathcal{C}_s(z) := \mathrm{sign}(z) \times \|z\|_2 \times \frac{\chi}{s}$. $\chi \in \mathbb{R}^d$ is a random vector with j-th element defined as: $\chi := \begin{cases} l+1 & \text{with probability } s\frac{|z_j|}{\|z\|_2} - l\,, \\ l & \text{otherwise} \end{cases}$ where the level l is such that $\frac{s|z_j|}{\|z\|_2} \in [l, l+1[$.*

The s-quantization scheme verifies Assumption 5 with $\omega = \min(d/s^2, \sqrt{d}/s)$ [the proof can be found in 2, see Appendix A.1].

We first consider two simple synthetic datasets: one for least-squares regression (with the same distribution over each machine), and one for logistic regression (with varying distributions across clients).

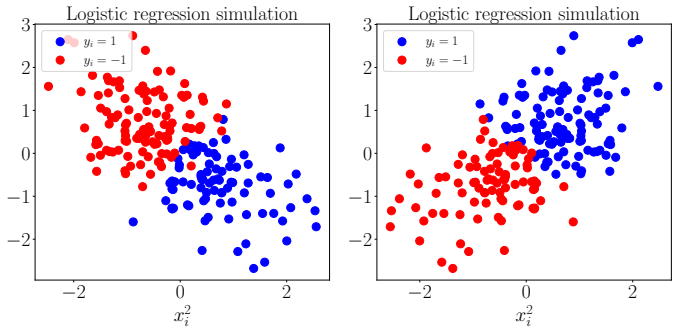
4.1. Synthetic datasets

The aim of using synthetic datasets is to underline the properties resulting from Theorems 1 to 3. We build two different synthetic datasets for i.i.d. or non-i.i.d. cases. We use linear regression to tackle the i.i.d case and logistic regression to handle the non-i.i.d. settings. Each worker i holds n_i observations $(z_j^i)_{1\leq j\leq n_i} = (x_j^i, y_j^i)_{1\leq j\leq n_i} = (X^i, Y^i)$ following a distribution D_i.

We use $N = 20$ devices, each holding 200 points of dimension $d = 20$ for least-square regression and $d = 2$ for logistic regression. We ran algorithms over 100 epochs.

Choice of the step-size for the synthetic datasets. For stochastic descent, we use a decreasing step-size $\gamma_k = \frac{1}{L\sqrt{k}}$ with k in $\mathbb{N}$, and for the full gradient descent we choose $\gamma = \frac{1}{L}$.

For i.i.d. setting, we use a linear regression model without bias. For each worker i, data points are generated from a normal distribution $(x_j^i)_{1\leq j\leq n_i} \sim \mathcal{N}(0, \Sigma)$. And



(a) Dataset 1 (b) Dataset 2

Figure 1: Data distribution for logistic regression to simulate non-i.i.d. data. Half of the device holds the first dataset, and the other half the second one.

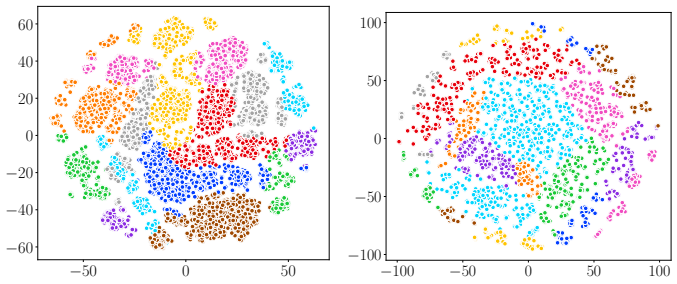
then, for all j in $[\![1, n_i]\!]$, we have: $y_j^i = \langle w, x_j^i \rangle + e_i$ with $e_i \sim \mathcal{N}(0, \lambda^2)$ and w the true model.

To obtain $\sigma_* = 0$, it is enough to remove the noise e_i by setting the variance λ^2 of the dataset distribution to 0. Indeed, using a least-square regression, for all i in $[\![1, N]\!]$, the cost function evaluated at point w is $F_i(w) = \frac{1}{2}\|X^{i^T} w - Y^i\|^2$. Thus the stochastic gradient j in $[\![1, n_i]\!]$ on device i in $[\![1, N]\!]$ is $g_j^i(w) = (X_j^{i^T} w - Y_j^i) X_j^i$. On the other hand, the true gradient is $\nabla F_i(w) = \mathbb{E} X^i X^{i^T}(w - w^*)$. Computing the difference, we have for all device i in $[\![1, N]\!]$ and all j in $[\![1, n_i]\!]$:

$$g_j^i(w) - F_i(w) = \underbrace{(X_j^i X_j^{i^T} - \mathbb{E} X^i X^{i^T})(w - w_*)}_{\text{multiplicative noise equal to 0 in } w_*} + \underbrace{(X_j^{i^T} w_* - Y_j^i)}_{\sim \mathcal{N}(0,\lambda^2)} X_j^i \tag{4}$$

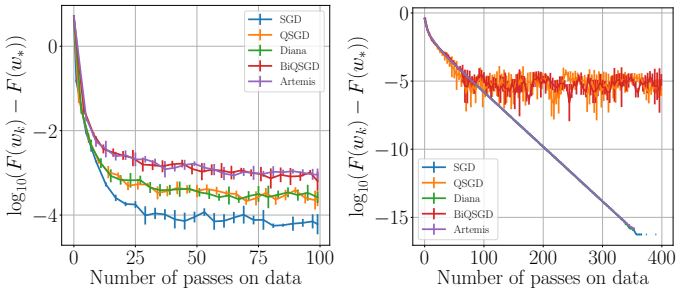
This is why, if we set $\lambda = 0$ and evaluate Equation (4) at w_*, we get back Assumption 3 with $\sigma_* = 0$, and as a consequence, the stochastic noise at the optimum is removed. Remark that it remains a stochastic gradient descent, and the uniform bound on the gradients noise **is not 0**. We set $\lambda^2 = 0 (\Leftrightarrow \sigma_*^2 = 0)$ in Figure S6. Otherwise, we set $\lambda^2 = 0.4$.

For non-i.i.d. setting, we generate two different datasets based on a logistic model with two different parameters: $w_1 = (10, 10)$ and $w_2 = (10, -10)$. Thus the model is expected to converge to $w_* = (10, 0)$. We have two different data distributions $x_1 \sim \mathcal{N}(0, \Sigma_1)$ and $x_2 \sim \mathcal{N}(0, \Sigma_2)$, and for all i in $[\![1, N]\!]$, for all k in $[\![1, n_i]\!]$, $y_k^i = \mathcal{R}\left(\text{Sigm}\left(\left\langle w_{(i \mod 2)+1}, x_{(i \mod 2)+1}^k \right\rangle\right)\right) \in \{-1, +1\}$. That is, half the machines use the first distribution $\mathcal{N}(0, \Sigma_1)$ for inputs and model w_1 and the other half the second distribution for inputs and model w_2. Here, $\mathcal{R}$ is the Rademacher distribution and Sigm is the sigmoid function defined as Sigm: $x \mapsto \frac{e^x}{1 + e^x}$. These two distributions are presented on Figure 1.



(a) Quantum dataset. (b) Superconduct dataset.

Figure 2: TSNE representations.



(a) LSR (i.i.d.): $\sigma_*^2 \neq 0$ (b) LR (non-i.i.d.): $\sigma_* = 0$.

Figure 3: Left: illustration of the saturation when $\sigma_* \neq 0$ and data is i.i.d., right: illustration of the memory benefits when $\sigma_* = 0$ but with non-i.i.d. data.

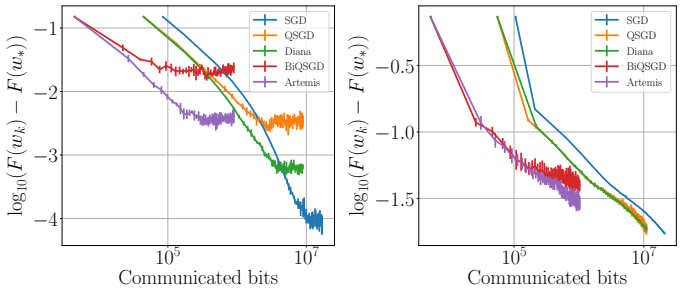
4.2. Real dataset

To illustrate theorems on real data and higher dimension, we then consider two real-world dataset: *superconduct* [see 11, with 21 263 points and 81 features] and *quantum* [see 6, with 50 000 points and 65 features] with $N = 20$ workers. To simulate non-i.i.d. *and* unbalanced workers, we split the dataset in heterogeneous groups, using a Gaussian mixture clustering on the TSNE representations (defined by Maaten & Hinton [24]). Thus, the data are highly non-i.i.d. *and* unbalanced over devices. We plot on Figure 2 the TSNE representation of the two real datasets. For *superconduct*, there are between 250 and 3900 points by worker, with a median at 750 ; and for *quantum*, there are between 900 and 10500 points, with a median at 2300.

4.3. Analysis of convergence

Convergence. Figure 3a presents the convergence of each algorithm w.r.t. the number of iterations k. During first iterations all algorithms make fast progress. However, because $\sigma_*^2 \neq 0$, all algorithms saturate; and the saturation level is higher for double compression (`Artemis`, `Bi-QSGD`), than for simple compression (`Diana`, `QSGD`), or than for SGD. This corroborates findings in Theorem 1 and Theorem 3.

Complexity. On Figure 4, the loss is plotted w.r.t. the theoretical number of bits exchanged after k iterations for the *quantum* and *superconduct* dataset. This confirms that double compression should be the method of choice to achieve a reasonable precision (*w.r.t.* σ_*), whereas for



(a) Quantum (b) Superconduct

Figure 4: **Real dataset** (non-i.i.d.): $\sigma_* \neq 0$, $N = 20$ workers, $p = 1$, $b > 1$ (150 iter.). X-axis in # bits.

high precision, a simple method like SGD results *in a lower complexity.*

Linear convergence under *null variance at the optimum.* To highlight the significance of our new condition on the noise, we compare $\sigma_*^2 \neq 0$ and $\sigma_*^2 = 0$ on Figure 3. Saturation is observed in Figure 3a, but if we consider a situation in which $\sigma_*^2 = 0$, and where the uniform bound on the gradient's variance *is not null* (as opposed to experiments in Liu et al. [23] who consider batch gradient descent), *a linear convergence rate is observed.* This illustrates that our new condition is sufficient to reach a linear convergence. Comparing Figure 3a with Figure S6a sheds light on the fact that the saturation level (before which double compression is indeed beneficial) is truly proportional to the noise variance *at optimal point* i.e. σ_*^2. And when $\sigma_*^2 = 0$, bidirectional compression is much more effective than the other methods (see Figure S6 in Appendix C.1).

Heterogeneity and real datasets. While in Figure 3a, data is i.i.d. on machines, and `Artemis` is thus not expected to outperform `Bi-QSGD` (the difference between the two being the memory), in Figures 3b and 4 we use **non-i.i.d. data**. None of the previous papers on compression directly illustrated the impact of heterogeneity on simple examples, neither compared it with i.i.d. situations.

5. Conclusion

We propose `Artemis`, a framework using bidirectional compression to reduce the number of bits needed to perform distributed or federated learning. On top of compression, `Artemis` includes a memory mechanism which improves convergence over non-i.i.d. data. We provide three tight theorems giving guarantees of a fast convergence (linear up to a threshold), highlighting the impact of memory, analyzing Polyak-Ruppert averaging and obtaining lowers bound by studying convergence in distribution of our algorithm. Altogether, this improves the understanding of compression combined with a memory mechanism and sheds light on challenges ahead.

Acknowledgments

We would like to thank Richard Vidal, Laeticia Kameni from Accenture Labs (Sophia Antipolis, France) and Eric Moulines from École Polytechnique for interesting discussions. This research was supported by the *SCAI: Statistics and Computation for AI* ANR Chair of research and teaching in artificial intelligence and by *Accenture Labs* (Sophia Antipolis, France).

Bibliography

[1] Agarwal, N., Suresh, A. T., Yu, F. X. X., Kumar, S., and McMahan, B. cpSGD: Communication-efficient and differentially-private distributed SGD. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 7564–7575. Curran Associates, Inc., 2018.

[2] Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding. *Advances in Neural Information Processing Systems*, 30:1709–1720, 2017.

[3] Alistarh, D., Hoefler, T., Johansson, M., Konstantinov, N., Khirirat, S., and Renggli, C. The Convergence of Sparsified Gradient Methods. *Advances in Neural Information Processing Systems*, 31:5973–5983, 2018.

[4] Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32): 15849–15854, 2019. Publisher: National Acad Sciences.

[5] Bottou, L. Online learning and stochastic approximations. 1999. doi: 10.1017/CBO9780511569920.003.

[6] Caruana, R., Joachims, T., and Backstrom, L. KDD-Cup 2004: results and analysis. *ACM SIGKDD Explorations Newsletter*, 6(2):95–108, December 2004. ISSN 1931-0145. doi: 10.1145/1046456.1046470.

[7] Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M., Senior, A., Tucker, P., Yang, K., Le, Q., and Ng, A. Large Scale Distributed Deep Networks. *Advances in Neural Information Processing Systems*, 25, 2012.

[8] Dieuleveut, A., Durmus, A., and Bach, F. Bridging the gap between constant step size stochastic gradient descent and markov chains. *Ann. Statist.*, 48(3):1348–1382, 06 2020. doi: 10.1214/19-AOS1850. URL https://doi.org/10.1214/19-AOS1850.

[9] Elias, P. Universal codeword sets and representations of the integers, September 1975.

[10] Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. SGD: General Analysis and Improved Rates. In *International Conference on Machine Learning*, pp. 5200–5209. PMLR, May 2019. ISSN: 2640-3498.

[11] Hamidieh, K. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354, November 2018. ISSN 0927-0256. doi: 10.1016/j.commatsci.2018.07.052.

[12] Horváth, S., Kovalev, D., Mishchenko, K., Richtárik, P., and Stich, S. Stochastic distributed learning with gradient quantization and double-variance reduction. *Optimization Methods and Software*, pp. 1–16, 2022.

[13] Horváth, S. and Richtárik, P. A Better Alternative to Error Feedback for Communication-Efficient Distributed Learning. *arXiv:2006.11077 [cs, stat]*, June 2020. arXiv: 2006.11077.

[14] Huang, Y., Cheng, Y., Bapna, A., Firat, O., Chen, D., Chen, M., Lee, H., Ngiam, J., Le, Q. V., Wu, Y., and Chen, z. GPipe:

Efficient Training of Giant Neural Networks using Pipeline Parallelism. In Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. d., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[15] Index, S. G. Speedtest Global Index – Monthly comparisons of internet speeds from around the world, 2020.

[16] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and Open Problems in Federated Learning. *arXiv:1912.04977 [cs, stat]*, December 2019. arXiv: 1912.04977.

[17] Karimireddy, S. P., Rebjock, Q., Stich, S., and Jaggi, M. Error Feedback Fixes SignSGD and other Gradient Compression Schemes. In *International Conference on Machine Learning*, pp. 3252–3261. PMLR, May 2019. ISSN: 2640-3498.

[18] Khirirat, S., Magnússon, S., Aytekin, A., and Johansson, M. Communication Efficient Sparsification for Large Scale Machine Learning. *arXiv:2003.06377 [math, stat]*, March 2020. arXiv: 2003.06377.

[19] Konečný, J., McMahan, H. B., Yu, F. X., Richtarik, P., Suresh, A. T., and Bacon, D. Federated Learning: Strategies for Improving Communication Efficiency. In *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

[20] Lannelongue, L., Grealey, J., and Inouye, M. Green algorithms: Quantifying the carbon footprint of computation. *Advanced Science*, pp. 2100707, 2021. Publisher: Wiley Online Library.

[21] Li, M., Andersen, D. G., Park, J. W., Smola, A. J., Ahmed, A., Josifovski, V., Long, J., Shekita, E. J., and Su, B.-Y. Scaling distributed machine learning with the parameter server. In *Proceedings of the 11th USENIX conference on Operating Systems Design and Implementation*, OSDI'14, pp. 583–598, USA, October 2014. USENIX Association. ISBN 978-1-931971-16-4.

[22] Li, Z., Kovalev, D., Qian, X., and Richtarik, P. Acceleration for Compressed Gradient Descent in Distributed and Federated Optimization. In *International Conference on Machine Learning*, pp. 5895–5904. PMLR, November 2020. ISSN: 2640-3498.

[23] Liu, X., Li, Y., Tang, J., and Yan, M. A Double Residual Compression Algorithm for Efficient Distributed Learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 133–143, June 2020. ISSN: 1938-7228 Section: Machine Learning.

[24] Maaten, L. v. d. and Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. ISSN ISSN 1533-7928.

[25] Mayekar, P. and Tyagi, H. RATQ: A Universal Fixed-Length Quantizer for Stochastic Optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1399–1409. PMLR, June 2020. ISSN: 2640-3498.

[26] McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, April 2017. ISSN: 2640-3498.

[27] Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv:1908.05355 [math, stat]*, October 2019. arXiv: 1908.05355.

[28] Meyn, S. and Tweedie, R. *Markov chains and stochastic stability*. Cambridge University Press, New York, NY, USA, 2 edition, 2009. ISBN 0-521-73182-8 978-0-521-73182-9.

[29] Mishchenko, K., Gorbunov, E., Takáč, M., and Richtárik, P. Distributed Learning with Compressed Gradient Differences. *arXiv:1901.09269 [cs, math, stat]*, June 2019. arXiv: 1901.09269.

[30] Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer US, 2004. ISBN 978-1-4020-7553-7. doi: 10.1007/978-1-4419-8853-9.

[31] Rabbat, M. G. and Nowak, R. D. Quantized incremental algorithms for distributed optimization. *IEEE Journal on Selected Areas in Communications*, 23(4):798–808, 2005.

[32] Rahimi, A. and Recht, B. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

[33] Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. *ICML*, 2012.

[34] Reisizadeh, A., Mokhtari, A., Hassani, H., Jadbabaie, A., and Pedarsani, R. FedPAQ: A Communication-Efficient Federated Learning Method with Periodic Averaging and Quantization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2021–2031. PMLR, June 2020. ISSN: 2640-3498.

[35] Robbins, H. and Monro, S. A Stochastic Approximation Method. *Annals of Mathematical Statistics*, 22(3):400–407, September 1951. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177729586. Number: 3 Publisher: Institute of Mathematical Statistics.

[36] Sattler, F., Wiedemann, S., Müller, K.-R., and Samek, W. Robust and Communication-Efficient Federated Learning From Non-i.i.d. Data. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2019. ISSN 2162-2388. doi: 10.1109/TNNLS.2019.2944481. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

[37] Seide, F., Fu, H., Droppo, J., Li, G., and Yu, D. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Fifteenth Annual Conference of the International Speech Communication Association*. Citeseer, 2014.

[38] Stich, S. U. Local SGD Converges Fast and Communicates Little. *arXiv:1805.09767 [cs, math]*, May 2019. arXiv: 1805.09767.

[39] Stich, S. U., Cordonnier, J.-B., and Jaggi, M. Sparsified SGD with Memory. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 4447–4458. Curran Associates, Inc., 2018.

[40] Strom, N. Scalable distributed DNN training using commodity GPU cloud computing. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[41] Tang, H., Yu, C., Lian, X., Zhang, T., and Liu, J. DoubleSqueeze: Parallel Stochastic Gradient Descent with Double-pass Error-Compensated Compression. In *International Conference on Machine Learning*, pp. 6155–6165. PMLR, May 2019. ISSN: 2640-3498.

[42] Vempala, S. S. *The random projection method*, volume 65. American Mathematical Soc., 2005.

[43] Villani, C. *Optimal transport : old and new*. Grundlehren der mathematischen wissenschaften. Springer, Berlin, 2009. ISBN 978-3-540-71049-3.

[44] Wen, W., Xu, C., Yan, F., Wu, C., Wang, Y., Chen, Y., and

Li, H. TernGrad: Ternary Gradients to Reduce Communication in Distributed Deep Learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 1509–1519. Curran Associates, Inc., 2017.

[45] Wu, J., Huang, W., Huang, J., and Zhang, T. Error Compensated Quantized SGD and its Applications to Large-scale Distributed Optimization. In *International Conference on Machine Learning*, pp. 5325–5333. PMLR, July 2018. ISSN: 2640-3498.

[46] Yu, Y., Wu, J., and Huang, L. Double Quantization for Communication-Efficient Distributed Optimization. In Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. d., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 4438–4449. Curran Associates, Inc., 2019.

[47] Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *5th international conference on learning representations, ICLR 2017, toulon, france, april 24-26, 2017, conference track proceedings*. OpenReview.net, 2017.

[48] Zheng, S., Huang, Z., and Kwok, J. Communication-Efficient Distributed Blockwise Momentum SGD with Error-Feedback. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[49] Zhou, S., Wu, Y., Ni, Z., Zhou, X., Wen, H., and Zou, Y. DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients. *arXiv:1606.06160 [cs]*, February 2018. arXiv: 1606.06160.

[50] Zhu, D. L. and Marcotte, P. Co-Coercivity and Its Role In the Convergence of Iterative Schemes For Solving Variational Inequalities, March 1996.

Bidirectional compression in heterogeneous settings for distributed or federated learning: tight convergence guarantees.

Supplementary material

In this appendix, we provide additional details to our work. In Appendix A, we define the filtrations used in our demonstrations. Secondly, in Appendix B, we analyze at a finer level the bandwidth speeds across the world to get a better intuition of the state of the worldwide internet usage. Thirdly, in Appendix C, we present the detailed framework of our experiments and give further illustrations to our theorems. In Appendix D, we gather a few technical results and introduce the lemmas required in the proofs of the main results. Those proofs are finally given in Appendix E. More precisely, Theorem 1 follows from Theorems S4 and S5, which are proved in Appendices E.1 and E.2, while Theorems 2 and 3 are respectively proved in Appendices E.3 and E.4.

Contents

Appendix A. Filtrations

In this section, we provide some explanations about filtrations - especially a rigorous definition - and how it is used in the proofs of Theorems 1 to 3.

Let a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with Ω a sample space, $\mathcal{A}$ a σ-algebra, and $\mathbb{P}$ a probability measure. We recall that the σ-algebra generated by a random variable $X : \Omega \to \mathbb{R}^m$ is

$$\sigma(X) = \{X^{-1}(A) : A \in \mathcal{B}(\mathbb{R}^m)\}\,,$$

where $\mathcal{B}(\mathbb{R}^m)$ is the Borel set of $\mathbb{R}^m$.

Furthermore, we recall that a filtration of $(\Omega, \mathcal{A}, \mathbb{P})$ is defined as an increasing sequence $(\mathcal{F}_n)_{n\in\mathbb{N}}$ of σ-algebras:

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}\,.$$

Randomness in our algorithm comes from three sources, therefore, we define three sequences of i.i.d. zero-centered random fields $(\xi_k^i)_{k\in\mathbb{N},i\in\{1,\dots,N\}}$, $(\epsilon_k^i)_{k\in\mathbb{N},i\in\{1,\dots,N\}}$, $(\epsilon_k)_{k\in\mathbb{N}}$.

1. Stochastic gradients. It corresponds to the noise associated with the computation of the stochastic gradient on device i at epoch k. We have:

$$\forall k \in \mathbb{N}^*\,,\ \forall i \in [\![0, ..., N]\!], \quad g_k^i = \nabla F_i(w_{k-1}) + \xi_k^i(w_{k-1})\,.$$

$$w_{k-1} \xrightarrow{\xi_k^i} g_k^i \xrightarrow{\epsilon_k^i} \widehat{g}_k^i \longrightarrow \widehat{g}_k = \sum_{i=1}^{N} \widehat{g}_k^i \xrightarrow{\epsilon_k} \Omega_k = \mathcal{C}\left(\widehat{g}_k\right)$$

Figure S1: The sequence of successive noises in the algorithm.

2. Uplink compression: this noise corresponds to the uplink compression when local gradients are compressed. Let $k \in \mathbb{N}$ and $i \in [\![0, ..., N]\!]$, suppose, we want to compress $\Delta_k^i \in \mathbb{R}^d$, then:

$$\forall k \in \mathbb{N}^*,\ \forall i \in [\![0, ..., N]\!], \quad \widehat{\Delta}_k^i = \Delta_k^i + \epsilon_k^i(\Delta_k^i) \iff \widehat{g}_k^i = g_k^i + \epsilon_k^i(\Delta_k^i)\,.$$

3. Downlink compression. This noise corresponds to the downlink compression when the global model parameter is compressed. Let $k \in \mathbb{N}$, suppose we want to compress $\widehat{g}_k \in \mathbb{R}^d$, then:

$$\forall k \in \mathbb{N}^*, \quad \Omega_k = \mathcal{C}_s(\widehat{g}_k) = \widehat{g}_k + \epsilon_k(\widehat{g}_k)\,.$$

This "succession of noises" in the algorithm is illustrated in Figure S1. In order to handle these three sources of randomness, we define three sequences of nested σ-algebras.

Definition 2. *We note $(\mathcal{F}_k)_{k\in\mathbb{N}}$ the filtration associated to the stochastic gradient computation noise, $(\mathcal{G}_k)_{k\in\mathbb{N}}$ the filtration associated to the uplink compression noise and $(\mathcal{H}_k)_{k\in\mathbb{N}}$ the filtration associated to the downlink compression noise. For $k \in \mathbb{N}^*$, we define:*

$$\begin{aligned}
\mathcal{F}_k &= \sigma\left(\Gamma_{k-1}, (\xi_k^i)_{i=1}^N\right)\\
\mathcal{G}_k &= \sigma\left(\Gamma_{k-1}, (\xi_k^i)_{i=1}^N, (\epsilon_k^i)_{i=1}^N\right)\\
\mathcal{H}_k &= \sigma\left(\Gamma_{k-1}, (\xi_k^i)_{i=1}^N, (\epsilon_k^i)_{i=1}^N, \epsilon_k\right)
\end{aligned}$$

with $\Gamma_k = \{(\xi_t^i)_{i\in[\![1,N]\!]}, (\epsilon_t^i)_{i\in[\![1,N]\!]}, \epsilon_t\}_{t\in[\![1,k]\!]}$ and $\Gamma_0 = \{\varnothing\}$.

We can make the following observations for all $k \geq 1$:

- From these three definitions, it follows that our sequences are nested.

$$\mathcal{F}_1 \subset \mathcal{G}_1 \subset \mathcal{H}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{H}_k\,.$$

- w_{k-1} is $\mathcal{H}_{k-1}$-measurable.
- g_k is $\mathcal{F}_k$-measurable.
- $\widehat{g}_k$ is $\mathcal{G}_k$-measurable.

As a consequence, we have Propositions S1 to S5. Below Proposition S1 gives the expectation over stochastic gradients conditionally to σ-algebras $\mathcal{H}_{k-1}$ and $\mathcal{F}_k$.

Proposition S1 (Stochastic Expectation)**.** *Let $k \in \mathbb{N}^*$ and $i \in [\![1, N]\!]$. Then on each local device $i \in [\![1, N]\!]$ we have almost surely (a.s.) $\mathbb{E}[g_k^i \mid \mathcal{F}_k] = g_k^i$ and $\mathbb{E}[g_k^i \mid \mathcal{H}_{k-1}] = \nabla F_i(w_{k-1})$.*

Proposition S2 gives expectation of uplink compression (information sent from remote devices to central server) conditionally to σ-algebras $\mathcal{F}_k$ and $\mathcal{G}_k$.

Proposition S2 (Uplink Compression Expectation)**.** *Let $k \in \mathbb{N}^*$ and $i \in [\![1, N]\!]$. Recall that $\widehat{g}_k^i = g_k^i + \epsilon_k^i$, then on each local device $i \in [\![1, N]\!]$, we have a.s. $\mathbb{E}[\widehat{g}_k^i \mid \mathcal{G}_k] = \widehat{g}_k^i$ and $\mathbb{E}[\widehat{g}_k^i \mid \mathcal{F}_k] = g_k^i$.*

From Assumption 5, it follows that variance over uplink compression can be bounded as expressed in Proposition S3.

Proposition S3 (Uplink Compression Variance)**.** *Let $k \in \mathbb{N}^*$ and $i \in [\![1, N]\!]$. Recall that $\Delta_k^i = g_k^i + h_{k-1}^i$, using Assumption 5 following hold a.s.:*

$$\mathbb{E}\left[\|\widehat{\Delta}_k^i - \Delta_k^i\|^2 \;\middle|\; \mathcal{F}_k\right] \leq \omega_{\mathrm{up}}\|\Delta_k^i\|^2 \tag{S1}$$

$$(\iff \quad \mathbb{E}\left[\|\widehat{g}_k^i - g_k^i\|^2 \;\middle|\; \mathcal{F}_k\right] \leq \omega_{\mathrm{up}}\|g_k^i\|^2 \text{ when no memory })\,. \tag{S2}$$

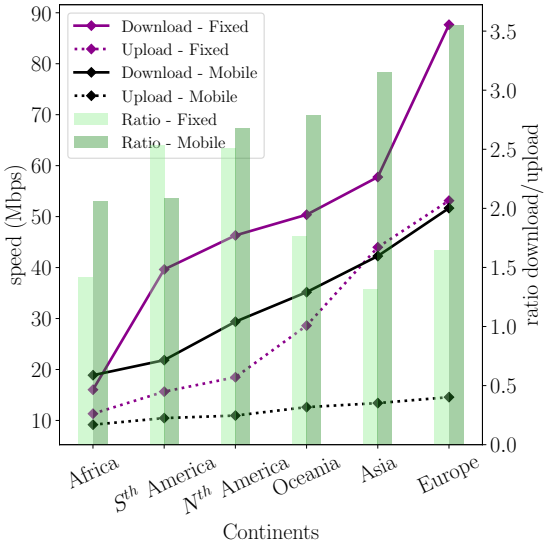


Figure S2: Left axis: upload and download speed for mobile and fixed broadband. Left axis: speeds (in Mbps), right axis: ratio (green bars). The dataset is gathered from *Speedtest.net*, see [15].

Concerning downlink compression (information sent from central server to each node), Proposition S4 gives its expectation w.r.t σ-algebras $\mathcal{G}_k$ and $\mathcal{H}_k$.

Proposition S4 (Downlink Compression Expectation)**.** *Let $k \in \mathbb{N}^*$, recall that $\Omega_k = \mathcal{C}_{\text{dwn}}(\widehat{g}_k) = \widehat{g}_k + \epsilon_k$, then a.s. $\mathbb{E}[\Omega_k \mid \mathcal{H}_k] = \Omega_k$ and $\mathbb{E}[\Omega_k \mid \mathcal{G}_k] = \widehat{g}_k$.*

The next proposition states that downlink compression can be bounded as for Proposition S3.

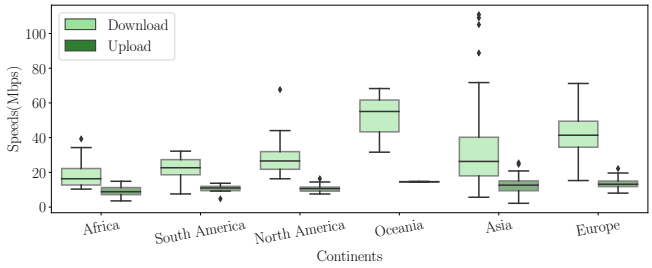
Proposition S5 (Downlink Compression Variance)**.** *Let $k \in \mathbb{N}$, using Assumption 5, we have a.s. $\mathbb{E}\left[\|\Omega_k - \widehat{g}_k\|^2 \;\middle|\; \mathcal{G}_k\right] \leq \omega_{\text{dwn}} \|\widehat{g}_k\|^2$.*

Appendix B. Bandwidth speed

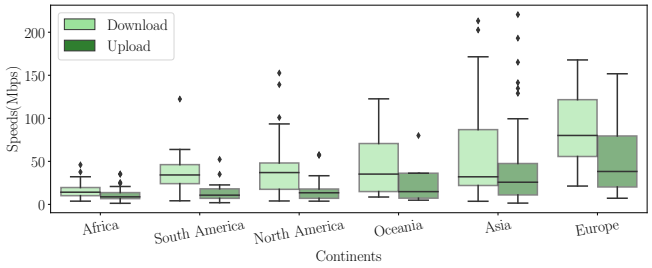
In a network configuration where download would be much faster than upload, bidirectional compression would present no benefit over unidirectional, as downlink communications would have a negligible cost. However, this is not the case in practice: to assess this point, we gathered broadband speeds, for both download and upload communications, for fixed broadband (cable, T1, DSL ...) or mobile (cellphones, smartphones, tablets, laptops ...) from studies carried out in 2020 over the 6 continents by *Speedtest.net* [see 15]. Results are provided in Figure S2, comparing download and upload speeds. The ratios (averaged by continents) between upload and download speeds stand between 1 (in Asia, for fixed broadband) and 3.5 (in Europe, for mobile broadband): there is thus no apparent reason to simply disregard the downlink communication, and bi-directionnal compression is unavoidable to achieve substantial speedup. More precisely, if we denote v_d and v_u the speed of download and upload (in Mbits per second), we typically have $v_d = \rho v_u$, with $1 < \rho < 3.5$. Using quantization with $s = 1$ (Definition 1), for unidirectional compression, each iteration takes $O\left(\frac{Nd}{\rho v_u}\right)$ seconds, while for a bidirectional one it takes only $O\left(\frac{N\sqrt{d}\log(d)}{v_u}\right)$ seconds.

The dataset is pickled from a study carried out by *Speedtest.net* [see 15]. This study has measured the bandwidth speeds in 2020 accross the six continents. In order to get a better understanding of this dataset, we illustrate the speeds distribution on Figures S2, S3a, S3b and S4.

In Figures S3a, S3b and S4, unlike Figure S2, we do not aggregate data by countries of a same continents. This allows to analyse the speeds ratio between upload and download with the *proper* value of each countries. Looking at Figures S3a, S3b and S4, it is noticeable that in the world, the ratio between upload and download speed is between 1 and 5, and not between 1 and 3.5 as Figure S2 was suggesting since we were aggregating data by continents. There are only nine countries in the world having a ratio higher than 5. In Europe : Malta, Belgium and Montenegro. In Asia : South Korea. In North America : Canada, Saint Vincent and the Grenadines, Panama and Costa Rica. In Africa : Western Sahara. The highest ratio is 7.7 observed in Malta.



(a) Mobile broadband.



(b) Fixed broadband.

Figure S3: Upload/download speed (in Mbps). Best seen incolors.

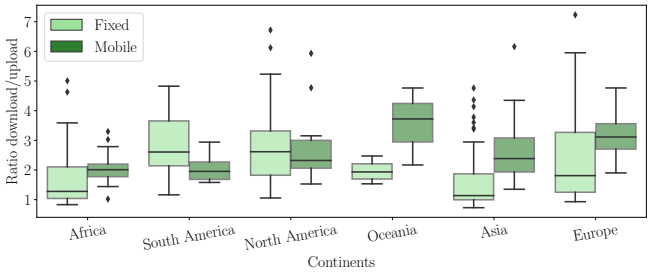


Figure S4: Distribution of the download/upload speeds ratio by continents. Best seen in colors.

Communication cost: an example using the quantization scheme. Using quantization (Definition 1), for any vector $v \in \mathbb{R}^d$, we are in possession of the tuple $(\|v\|^2, \phi, \chi)$, where ϕ is the vector of signs of $(v_j)_{j=1}^d$, and χ is the vector of integer values $(\chi_j)_{j=1}$. To broadcast the quantized value, we use the `Elias` encoding [9]. Using this encoding scheme, it can be shown (Theorem 3.2 of [2]) that:

Proposition S6. *For any vector v, the number of bits needed to communicate $\mathcal{C}_s(v)$ is upper bounded by:*

$$\left(3 + \left(\frac{3}{2} + o(1)\right) \log_2 \left(\frac{2(s^2+d)}{s(s+\sqrt{d})}\right)\right) s(s+\sqrt{d}) + 32\,.$$

With $s = 1$, it means that we will employ $O(\sqrt{d}\log_2 d)$ bits per iteration instead of $32d$, which reduces by a factor $\frac{\sqrt{d}}{\log_2 d}$ the number of bits used by iteration. Now, in a FL settings, at each iteration we have a double communication (device to the main server, main server to the device) for each of the N clients. It means that at each iteration, we need to communicate $2 \times N \times 32d$ bits if compression is not used. Obviously, unidirectional compression can at best result in a factor 2 reduction in term of total number of bits, while for bidirectional compression, we need to broadcast $O(N\sqrt{d}\log_2 d)$ bits using the `Elias` encoding [defined in 9]. Denoting v_d and v_u the speed of download and upload (in bits per second), we typically have $v_d = \rho v_u$, $3.5 > \rho > 1$. Then for unidirectional compression, each iteration takes $O\left(\frac{Nd}{v_d} + \frac{N\sqrt{d}\log_2(d)}{v_u}\right) = O\left(\frac{Nd}{\rho v_u}\right)$ seconds, while for a bidirectional one, it takes only $O\left(\frac{N\sqrt{d}\log_2(d)}{v_u}\right)$ seconds.

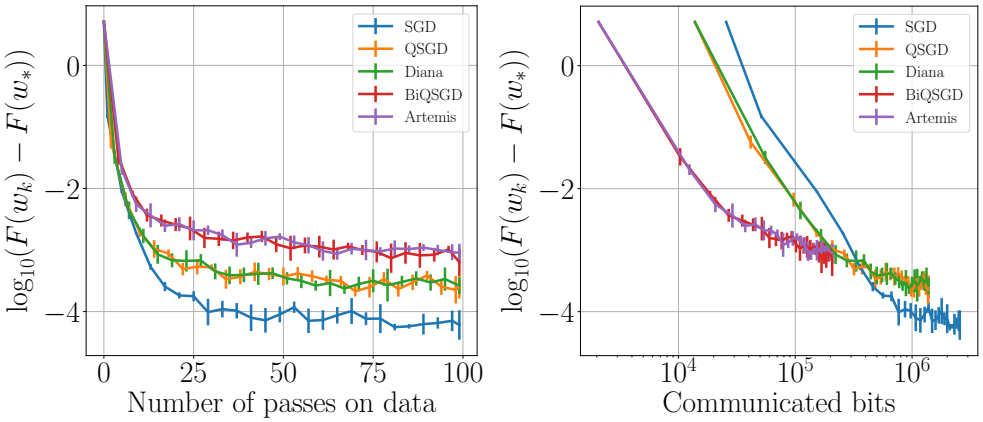
In other words, unless ρ is really large (which is not the case in practice as stressed by Figure S2), double compression reduces by several orders of magnitude the global time complexity, and bidirectional compression is superior to unidirectional.

Appendix C. Experiments

In this section we provide additional details about our experiments. We recall that we use two kind of datasets: 1) toy-ish synthetic datasets and 2) real datasets: *superconduct* [11, 21263 points, 81 features] and *quantum* [6, 50,000 points, 65 features]. The aim of using synthetic datasets is mainly to underline the properties resulting from Theorems 1 to 3.

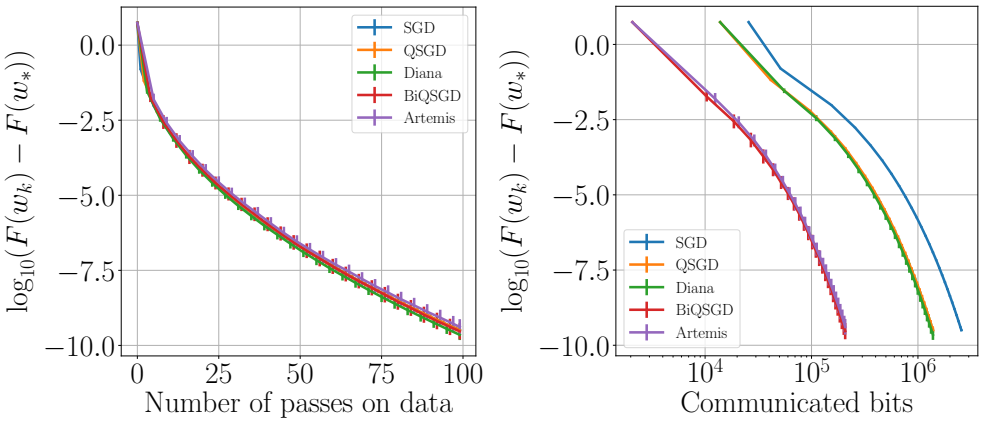
We use the same 1-quantization scheme (see Definition 1, $s = 1$ is the most drastic compression) for both uplink and downlink, and thus, we consider that $\omega_{\text{up}} = \omega_{\text{dwn}}$. In addition, we choose $\alpha_{\text{up}} = \frac{1}{2(1+\omega^{\text{up}})}$. All the figures can be found in the notebooks provided on our GitHub repository.

For each figure, we plot the convergence w.r.t. the number of iteration k or w.r.t. the theoretical number of bits exchanged after k iterations. On the Y-axis we display $\log_{10}(F(w_{k-1}) - F(w_*))$, with k in $\mathbb{N}$. All experiments have been



(a) LSR: $\sigma_*^2 \neq 0$ (b) X-axis in # bits.

Figure S5: **Synthetic dataset, Least-Square Regression with noise** ($\sigma_* \neq 0$). In a situation where data is i.i.d., the memory does not present much interest and has no impact on the convergence. Because $\sigma_*^2 \neq 0$, all algorithms saturate; and saturation level is higher for double compression (`Artemis`, `Bi-QSGD`), than for simple compression (`Diana`, `QSGD`) or than for SGD. This corroborates findings in Theorem 1 and Theorem 3.



(a) LSR: $\sigma_*^2 = 0$ (b) X-axis in # bits.

Figure S6: **Synthetic dataset, Least-Square Regression without noise** ($\sigma_* = 0$). Without surprise, with i.i.d data and $\sigma_* = 0$, the convergence of each algorithm is linear. Thus, in i.i.d. settings, the impact of the memory is negligible, but this will not be the case in the non-i.i.d. settings as underlined by Figure S7.

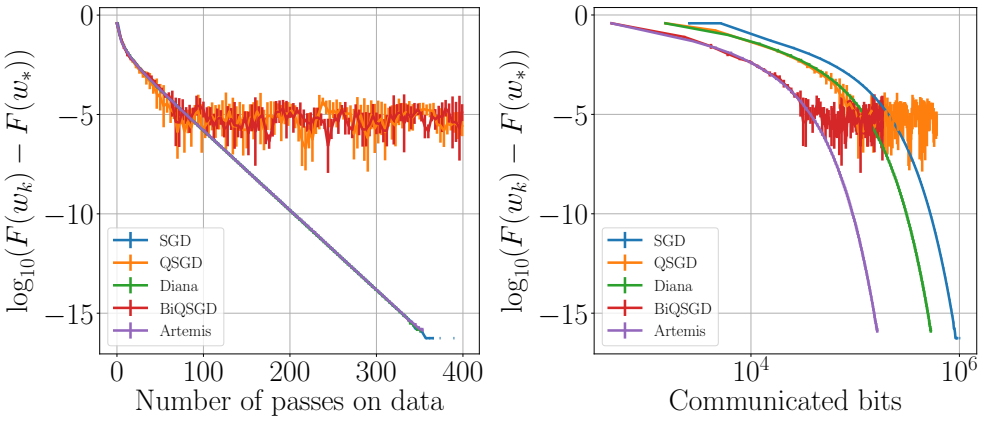
run 5 times and averaged before displaying the curves. We plot error bars on all figures. To compute error bars we take the standard deviation of $\log_{10}(F(w_{k-1}) - F(w_*))$, we then plot the curve $\pm$ this standard deviation.

Appendix C.1. Least-squares regression

In this Subsection, we present all figures generated using Least-squares regression. Note that Figure S5 corresponds to Figure 3a. As explained in the main body of the paper, in the case of $\sigma_* \neq 0$ (Figure S5), algorithms using memory (i.e `Diana` and `Artemis`) are not expected to outperform those without (i.e `QSQGD` and `Bi-QSGD`). On the contrary, they saturate at a higher level. However, as soon as the noise at the optimum is 0 (Figure S6), all algorithms (regardless of memory), converge at a linear rate exactly as classical SGD.

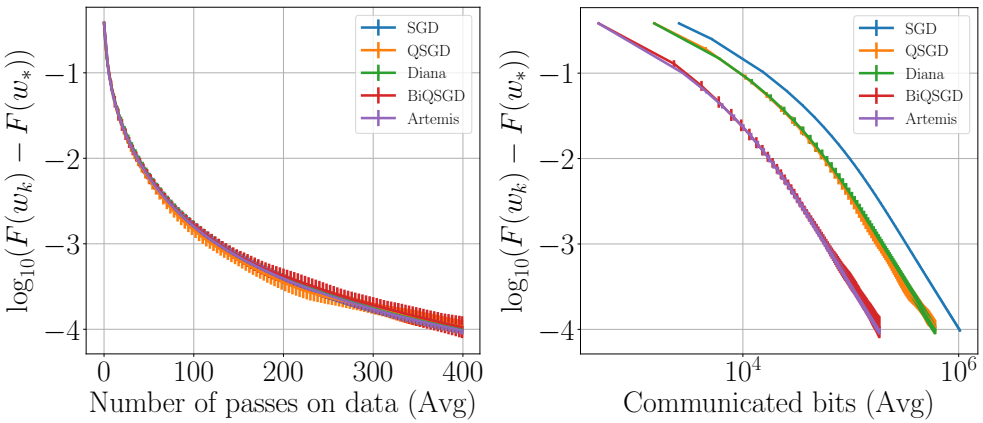
Appendix C.2. Logistic regression

In this Subsection, we present all figures generated using a logistic regression model. Note that Figure S7 corresponds to Figure 3b. Data is non-i.i.d. and we use a full batch gradient descent to get $\sigma_* = 0$ to shed light on the impact of memory on convergence. Figure S8 use the same setting than Figure S7 except that *it adds a Polyak-Ruppert averaging.* Note that in the absence of memory the variance increases compared to algorithms using memory. To generate these figures, we didn't take the optimal step-size. But if we took it, the trade-off between variance and bias would be worse and algorithms using memory would outperform those without.



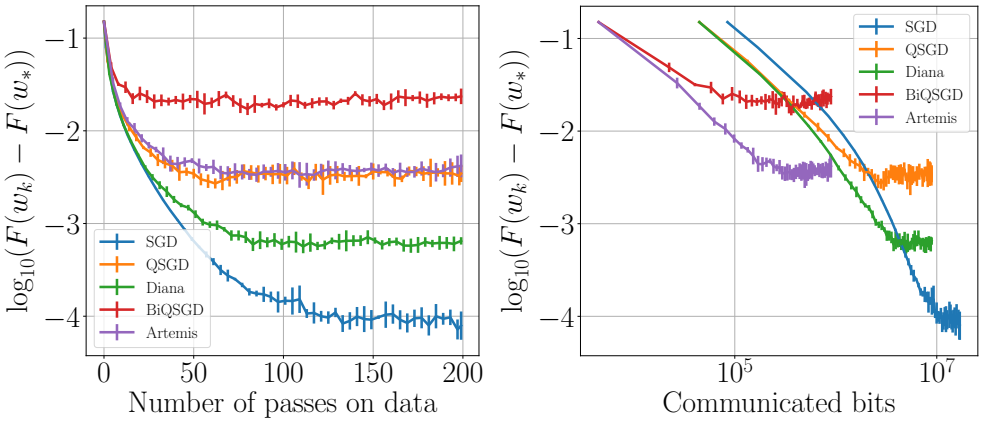
(a) LR: $\sigma_*^2 = 0$ (b) X-axis in # bits.

Figure S7: **Synthetic dataset, Logistic Regression on non-i.i.d. data** using a full batch gradient descent (to get $\sigma_* = 0$). The benefit of memory is obvious, it makes the algorithm converge linearly, while algorithms without are saturating at a higher level. This stresses the importance of using the memory in non-i.i.d. settings.



(a) LR: $\sigma_*^2 = 0$ (b) X-axis in # bits.

Figure S8: **Polyak-Ruppert averaging, synthetic dataset.** Logistic regression on non-i.i.d. data using a full batch gradient descent (to get $\sigma_* = 0$) and a Polyak-Ruppert averaging. The convergence is sublinear as predicted by Theorem 2 because $\sigma_* = 0$.



(a) X-axis in # epoch

(b) X-axis in # bit

Figure S10: ***Quantum***. Least-squares regression, $\sigma_* \neq 0$, $\gamma = 1/L$, $b = 256$, non-i.i.d..

Appendix C.3. Real datasets: Quantum and Superconduct

In this Subsection, we present details about experiments conducted on real-life datasets: *superconduct* (from Caruana et al. [6]) where we use least-squares regression, and *quantum* (from Hamidieh [11]) with logistic regression.

In the following, we present results on superconduct and quantum in the setting of full device participation. Next, we address in Appendix C.3.1 the issue of the optimal step-size. In Appendix C.3.2 we compare `Artemis` to other existing algorithms doing compression in a distributed learning framework. Finally, we estimate in Appendix C.4 the carbon footprint of the experiments presented in this article.

Convex settings are given in Figure S9. Experiments have been performed with 200 epochs in the stochastic regime, and 400 epochs in the full batch regime. We use quantization (Definition 1) with $s = 2^0$ for all experiments.

Figures S10 to S13 underline the benefit of using memory in the stochastic and full batch regime for non-i.i.d. datasets. Figures S10 and S12 correspond to Figure 4. We observe on these figures the benefit of the memory. The level of saturation of algorithms using memory is much lower than those without memory. Additionally, Theorem 1 highlights that the level of saturation (see constant E of Table 2) is proportional to the level of compression $\omega_{\text{up/dwn}}$. This is indeed observed on Figures S10 to S13.

In the case of the *quantum* dataset (see Figure S10), `Artemis` is not only better than `Bi-QSGD`, but in fact, as good as `QSGD`. That is to say, we achieve to make an algorithm doing bidirectional compression, as good as an algorithm doing unidirectional compression.

Figure S9: Settings of experiments.

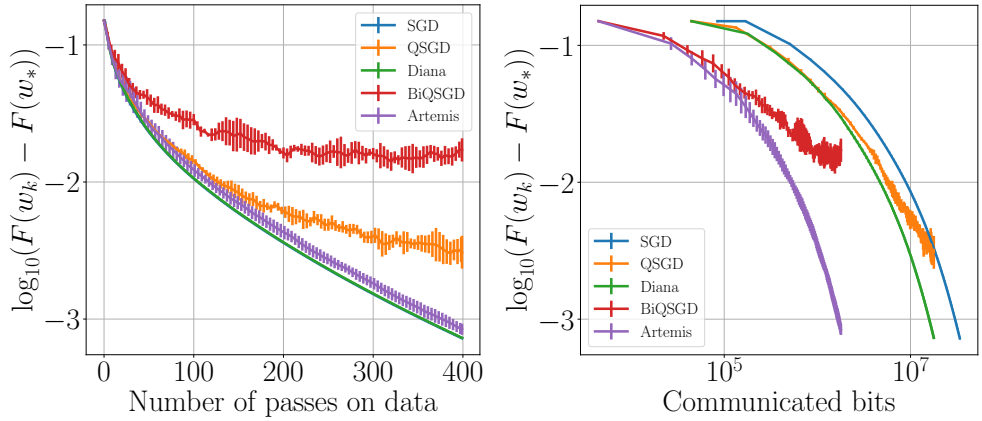
Settings	quantum	superconduct
references	[6]	[11]
model	LR	LSR
dimension d	66	82
training dataset size	$50,000$	$21,200$
batch size b	256	64
compression rate s	2^0 (*i.e.* two levels)	
norm quantization	$\|\cdot\|_2$	
momentum m	no momentum	
step-size γ	$1/L$	

On Figures S11 and S13, we run the five algorithms with full gradient descent, resulting in $\sigma_* = 0$. In this case, as the dependency on B^2 is removed, Theorem 1 predicts that we must have a linear convergence for algorithms using memory. This is experimentally observed.

Memory trade-off: batch size, noise at the optimum, and heterogeneity. Because the variance of the algorithm (see constant E of Table 2) is divided by the batch size b, the choice of this hyperparameter is not without importance. Indeed, reducing the batch size will increase the impact of σ_* on the convergence's rate, while the impact of B^2 will remain constant. Thus, there is a *trade-off*: if the batch-size is too small, the quantity σ_*/b will become larger than B^2, and the impact of the memory will be hidden by the second term depending on the dataset heterogeneity. This will lead `Artemis`-like algorithms to fail: the memory term is canceled by the high heterogeneity. On the other hand, if the dataset does not present enough heterogeneity, the constant B^2, will be negligible making memory useless, or even penalizing.

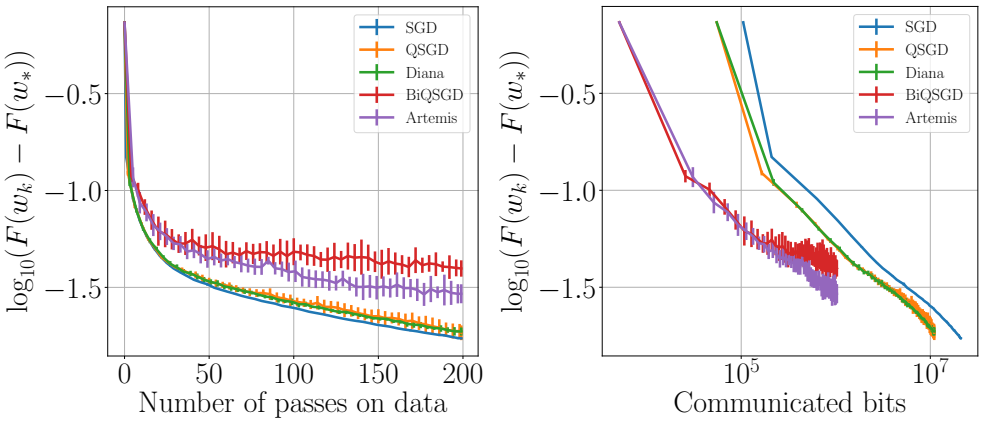
Appendix C.3.1. Optimized step-size

In this section, we want to address the issue of the optimal step-size. On Figure S14 we plot the minimal loss after 250 iterations for each of the 5 algorithms. We can see that algorithms with memory clearly outperform those without.



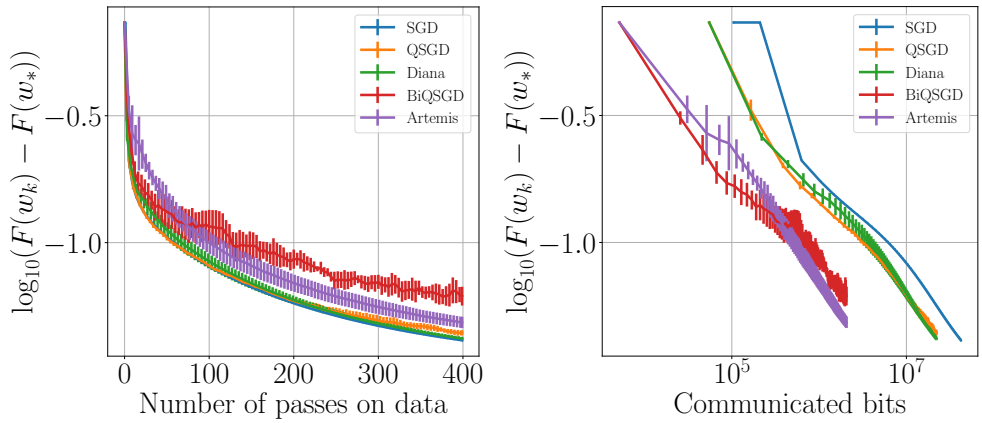
(a) X-axis in # epoch (b) X-axis in # bit

Figure S11: ***Quantum***. Least-squares regression, $\sigma_* = 0$, $\gamma = 1/L$, full gradient descent, non-i.i.d..



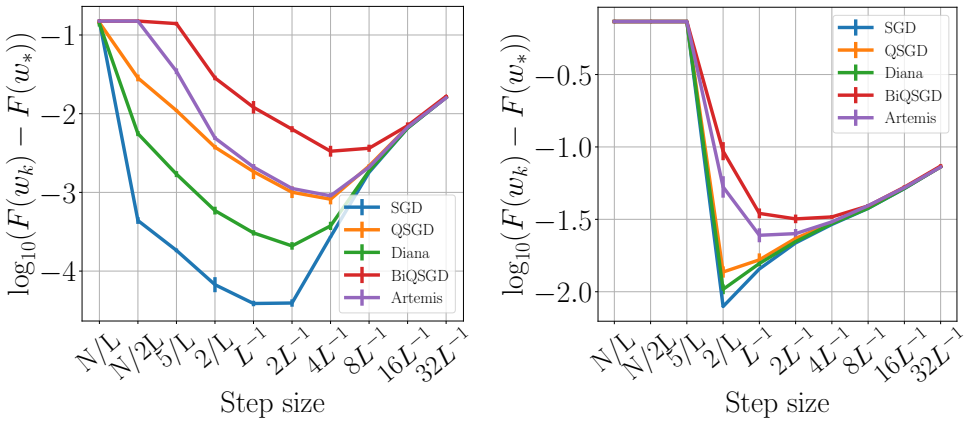
(a) X-axis in # epochs (b) X-axis in # bit

Figure S12: ***Superconduct***. Least-squares regression, $\sigma_* \neq 0$, $\gamma = 1/L$, $b = 64$, non-i.i.d..



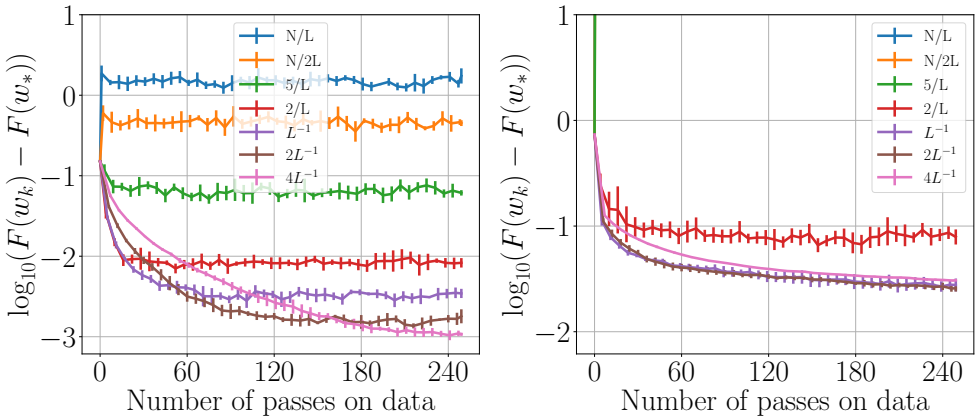
(a) X-axis in # epochs (b) X-axis in # bit

Figure S13: ***Superconduct***. Least-squares regression, $\sigma_* = 0$, $\gamma = 1/L$, full batch gradiend descent, non-i.i.d..



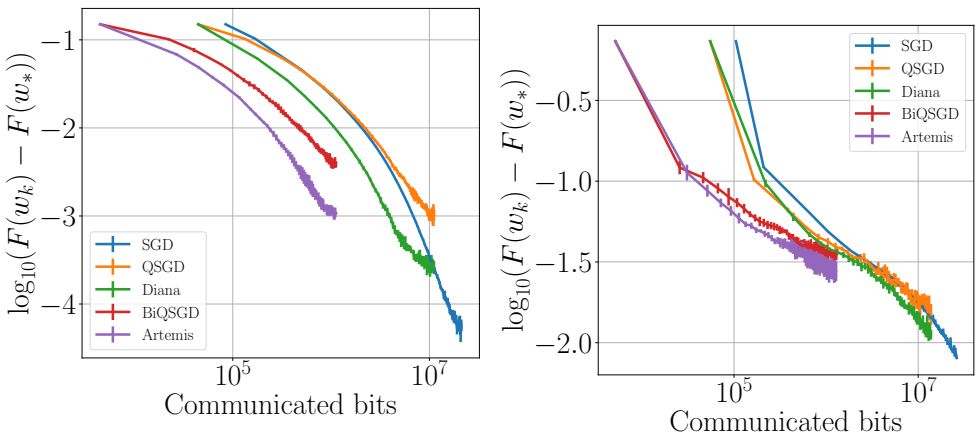
(a) Quantum (b) Superconduct

Figure S14: **Searching for the optimal step-size γ_{opt} for each algorithm.** X-axis - value on step-size, Y-axis - minimal loss after running 250 iterations



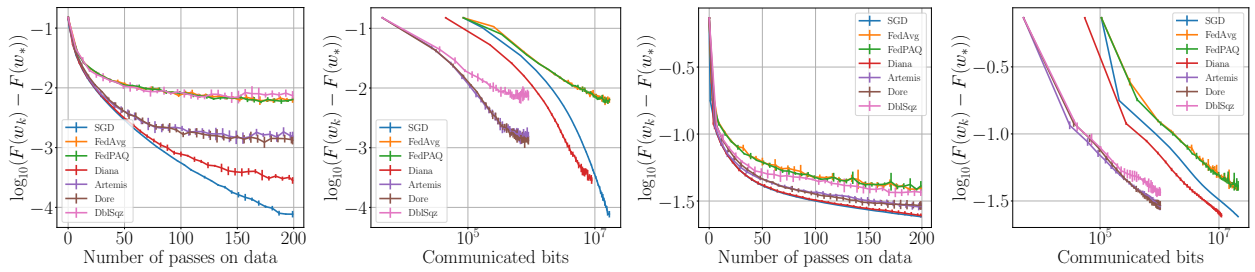
(a) Quantum (b) Superconduct

Figure S15: Loss w.r.t. step-size γ.



(a) Quantum (b) Superconduct

Figure S16: **Optimal step-size for each of the algorithms.** X-axis in # bits.



(a) Quantum (b) Quantum (c) Superconduct (d) Superconduct

Figure S17: **Artemis compared to other existing algorithms.** $\gamma = 1/(2L)$, X-axis in # epoch or in # bits.

Then, on Figure S15 we present the loss of `Artemis` after 250 iterations for various step-size: $\frac{N=20}{2L}$, $\frac{5}{L}$, $\frac{2}{L}$, $\frac{1}{L}$, $\frac{1}{2L}$, $\frac{1}{4L}$, $\frac{1}{8L}$ and $\frac{1}{16L}$. This helps to understand which step-size should be taken to obtain the best accuracy after k in $[\![1, 150]\!]$ iterations. Finally, on Figure S16, we plot the loss obtained with the optimal step-size γ_{opt} of each algorithms (found with Figure S14) w.r.t the number of communicated bits.

On Figure S14, it is interesting to note that the memory allows to increase the maximal step-size. So, the optimal step-size is $\gamma_{opt} = \frac{1}{L}$ for `Artemis` , but is $\gamma_{opt} = \frac{1}{2L}$ for `BiQSGD`.

We plot the loss of `Artemis` after 250 iterations for different step-size on Figure S15. As stressed by Figure S14, after 250 iterations, the best accuracy for both datasets is indeed obtained with $\gamma_{opt} = \frac{1}{L}$. And we observe that (as for Vanilla `SGD`), the optimal step-size of `Artemis` decreases with the number of iterations (e.g., for *quantum*, it is $1/L$ before 50 iterations and $1/2L$ after). This is consistent with Theorem 1.

Figure S16 plots the loss of each algorithm obtained with its optimal step-size γ i.e. the step-size that attains the lowest error after 150 iterations. For instance $\gamma = \frac{1}{L}$ for `Artemis`, but $\gamma = \frac{2}{L}$ for `SGD`. For both *superconduct* and *quantum* datasets, taking the optimal step-size leads `Artemis` to superior performance than other variants w.r.t. both accuracy and number of bits.

In conclusion of this subsection, Figures S14 to S16 allow to conclude on the significant impact of memory in a non-i.i.d. settings, and to claim that bidirectional compression with memory is by far superior (up to a threshold) to the four other algorithm: `SGD, QSGD, Diana` and `BiQSGD`.

Appendix C.3.2. Comparing `Artemis` with other existing algorithms

On Figure S17 we compare `Artemis` with other existing algorithms: `FedSGD`, `FedPAQ`, `Diana`, `Dore` and `Double-Squeeze`. We take $\gamma = 1/(2L)$ because otherwise `FedSGD` and `FedPAQ` diverge. These two algorithms present worse performance because they have not been designed for non-i.i.d. datasets.

We can observe that `Double-Squeeze` (which only uses error-feedback) is outperformed by `Artemis`. Besides, we observe that `Dore` (which combines this mechanism with memory) has identical rate of convergence than `Artemis`. It underlines that for unbiased operators of compression, **the enhancement comes from the memory and not from the error-feedback**.

`FedPAQ` (unidirectional compression) has a very fast convergence during first iterations, but then saturates at a level higher than for `Artemis`-like algorithms. `FedSGD` (no compression) presents a convergence's rate worse that vanilla `SGD` because it does not correctly handle heterogeneous datasets.

Appendix C.4. CPU usage and carbon footprint

As part as a community effort to report the amount of experiments that were performed, we estimated that overall our experiments ran for 220 to 270 hours end to end. We used an Intel(R) Xeon(R) CPU E5-2667 processor with 16 cores.

The carbon emissions caused by this work were subsequently evaluated with `Green Algorithm` built by Lannelongue et al. [20]. It estimates our computations to generate 30 to 35 kg of CO2, requiring 100 to 125 kWh. To compare, it corresponds to about 160 to 200km by car. This is a relatively moderate impact, matching the goal to keep the experiments for an illustrative purpose.

Appendix D. Technical results

In this section, we introduce a few technical lemmas that will be used in the proofs of Theorems S4 to S6. In Appendix D.1, we provide some classical results that are used throughout this article, in Appendix D.2 we present lemmas common to the proofs with/without memory and which are needed to prove the contraction of the Lyapunov function. Then, in respectively Appendices D.3 and D.4, we give lemmas adapted to the cases without and with memory.

Appendix D.1. Useful identities and inequalities

Lemma S1. *Let $N \in \mathbb{N}$ and $d \in \mathbb{N}$. For any sequence of vector $(a_i)_{i=1}^N \in \mathbb{R}^d$, we have the following inequalities:*

$$\left\| \sum_{i=1}^N a_i \right\|^2 \leq \left(\sum_{i=1}^N \|a_i\| \right)^2 \leq N \sum_{i=1}^N \|a_i\|^2 .$$

The first part of the inequality corresponds to the triangular inequality, while the second part is Cauchy's inequality.

Lemma S2. *Let $\alpha \in [0,1]$ and $x, y \in (\mathbb{R}^d)^2$, then:*

$$\|\alpha x + (1-\alpha)y\|^2 = \alpha \|x\|^2 + (1-\alpha)\|y\|^2 - \alpha(1-\alpha)\|x-y\|^2 .$$

This is a norm's decomposition of a convex combination.

Lemma S3. *Let X be a random vector of $\mathbb{R}^d$, then for any vector $x \in \mathbb{R}^d$:*

$$\mathbb{E}\|X - \mathbb{E}X\|^2 = \mathbb{E}\|X - x\|^2 - \|\mathbb{E}X - x\|^2 .$$

This equality is a generalization of the well know decomposition of the variance (with $x = 0$).

Lemma S4. *If $F : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ is strongly convex, then the following inequality holds:*

$$\forall (x,y) \in \mathbb{R}^d, \langle \nabla F(x) - \nabla F(y), x - y \rangle \geq \mu \|x-y\|^2 .$$

This inequality is a consequence of strong convexity and can be found in [30, equation 2.1.22].

Appendix D.2. Lemmas for proof of convergence

Below are presented technical lemmas needed to prove the contraction of the Lyapunov function for Theorems S4 and S5. In this section we assume that Assumptions 1 to 5 are verified. In Appendices D.3 and D.4 we separate lemmas that required only for the case with memory or without.

The first lemma is very simple and straightforward from the definition of Δ_k^i. We remind that Δ_k^i is the difference between the computed gradient and the memory hold on device i. It corresponds to the information which will be compressed and sent from device i to the central server.

Lemma S5 (Bounding the compressed term)**.** *The squared norm of $(\Delta_k^i)_{k\in\mathbb{N}^*, i\in\{1,\dots,N\}}$, the term sent by each node to the central server, can be bounded as follows:*

$$\forall k \in \mathbb{N}^*, \forall i \in [\![1, N]\!], \quad \left\|\Delta_k^i\right\|^2 \leq 2\left(\left\|g_k^i - h_*^i\right\|^2 + \left\|h_{k-1}^i - h_*^i\right\|^2 \right) .$$

Proof

Let $k \in \mathbb{N}$ and $i \in \{1, \dots, N\}$, we have by definition:

$$\left\|\Delta_k^i\right\|^2 = \left\|g_k^i - h_{k-1}^i\right\|^2 = \left\|(g_k^i - h_*^i) + (h_*^i - h_{k-1}^i)\right\|^2 .$$

Applying Lemma S1 gives the expected result. ■

Below, we show up a recursion over the memory term h_{k-1}^i involving the stochastic gradients. This recursion will be used in Lemma S12. This recursion has been first shed into light by Mishchenko et al. [29].

Lemma S6 (Expectation of memory term)**.** *The memory term h_k^i can be expressed using a recursion involving the stochastic gradient g_k^i:*

$$\forall k \in \mathbb{N}^*, \forall i \in [\![1, N]\!], \quad \mathbb{E}\left[h_k^i \;\middle|\; \mathcal{F}_k\right] = (1-\alpha_{\mathrm{up}})h_{k-1}^i + \alpha_{\mathrm{up}} g_k^i .$$

Proof Let $k \in \mathbb{N}$ and $i \in \{1, \dots, N\}$. We just need to decompose h_k^i using its definition:

$$h_k^i = h_{k-1}^i + \alpha_{\mathrm{up}}\widehat{\Delta}_k^i = h_{k-1}^i + \alpha_{\mathrm{up}}(\widehat{g}_k^i - h_{k-1}^i) = (1-\alpha_{\mathrm{up}})h_{k-1}^i + \alpha_{\mathrm{up}}\widehat{g}_k^i ,$$

and considering that $\mathbb{E}\left[\widehat{g}_k^i \;\middle|\; \mathcal{F}_k\right] = g_k^i$ (Proposition S2), the proof is completed. ■

In Lemma S7, we rewrite $\|g_k\|^2$ and $\left\|g_k - h_*^i\right\|^2$ to make appears:

1. the noise over stochasticity,
2. $\|g_k - g_{k,*}\|^2$ which is the term on which will later be applied cocoercivity (see Assumption 2).

Lemma S7 is required to correctly apply cocoercivity in Lemma S13.

Lemma S7 (Before using co-coercivity)**.** *Let $k \in [\![0, K]\!]$ and $i \in [\![1, N]\!]$. The noise on the stochastic gradients as defined in Assumptions 3 and 4 can be controlled as following:*

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_k^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \leq \frac{2}{N}\sum_{i=1}^{N}\left(\mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] + (\frac{\sigma_*^2}{b} + B^2)\right), \tag{S1}$$

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_k^i - h_*^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \leq \frac{2}{N}\sum_{i=1}^{N}\left(\mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] + \frac{\sigma_*^2}{b}\right). \tag{S2}$$

Proof Let $k \in \mathbb{N}$ and i in $\{1, \dots, N\}$. We obtain Equation (S1) using Lemma S1:

$$\left\|g_k^i\right\|^2 = \left\|g_k^i - g_{k,*}^i + g_{k,*}^i\right\|^2 \leq 2\left(\left\|g_k^i - g_{k,*}^i\right\|^2 + \left\|g_{k,*}^i\right\|^2\right).$$

Taking expectation with regards to filtration $\mathcal{H}_{k-1}$ and using Assumptions 3 and 4 gives the first result.

For Equation (S2), we use again Lemma S1 and we write (by definition, $h_*^i = \nabla F_i(w_*)$):

$$\left\|g_k^i - h_*^i\right\|^2 = \left\|\left(g_k^i - g_{k,*}^i\right) + \left(g_{k,*}^i - \nabla F_i(w_*)\right)\right\|^2 \leq 2(\left\|g_k^i - g_{k,*}^i\right\|^2 + \left\|g_{k,*}^i - \nabla F_i(w_*)\right\|^2).$$

Taking expectation, we have:

$$\begin{aligned}\mathbb{E}\left[\left\|g_k^i - h_*^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] &\leq 2\left(\mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] + \mathbb{E}\left[\left\|g_{k,*}^i - \nabla F_i(w_*)\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right]\right) \\ &\leq 2\left(\mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] + \frac{\sigma_*^2}{b}\right) \quad \text{using Assumption 3.}\end{aligned}$$

■

Demonstrating that the Lyapunov function is a contraction requires to bound $\|g_k\|^2$ which needs to control each term $(\|g_k^i\|^2)_{i=1}^N$ of the sum. This leads to invoke smoothness of F (consequence of Assumption 2).

Lemma S8. *Regardless if we use memory, we have the following bound on the squared norm of the gradient, for all k in $\mathbb{N}^*$:*

$$\mathbb{E}\left[\|g_k\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \leq \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_k^i - h_*^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] + L\left\langle\nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle.$$

Proof

Let $k \in \mathbb{N}^*$, taking expectation w.r.t the σ-algebra $\mathcal{H}_{k-1}$:

$$\mathbb{E}\left[\|g_k\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] = \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} g_k^i - \nabla F_i(w_{k-1}) + \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(w_{k-1})\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right].$$

Decomposing the squared norm:

$$\begin{aligned}\mathbb{E}\left[\|g_k\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] = \mathbb{E}&\left[\left\|\frac{1}{N}\sum_{i=1}^{N} g_k^i - \nabla F_i(w_{k-1})\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \\ &+ 2\mathbb{E}\left[\left\langle\frac{1}{N}\sum_{i=1}^{N} g_k^i - \nabla F_i(w_{k-1}), \frac{1}{N}\sum_{i=1}^{N}\nabla F_j(w_{k-1})\right\rangle \,\middle|\, \mathcal{H}_{k-1}\right] \\ &+ \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(w_{k-1})\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right].\end{aligned}$$

Moreover, $\forall i,j \in \{1,\dots,N\}^2, \mathbb{E}\left[\left\langle g_k^i - \nabla F_i(w_{k-1}), \nabla F_j(w_{k-1})\right\rangle \;\middle|\; \mathcal{H}_{k-1}\right] = 0$ and $\nabla F(w_{k-1})$ is $\mathcal{H}_{k-1}$-measurable, hence:

$$\mathbb{E}\left[\left\|g_k\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N g_k^i - \nabla F_i(w_{k-1})\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] + \left\|\nabla F(w_{k-1})\right\|^2 . \tag{S3}$$

To compute $\left\|\nabla F(w_{k-1})\right\|^2$, we apply cocoercivity (Assumption 2):

$$\left\|\nabla F(w_{k-1})\right\|^2 \leq L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle .$$

We note $\square_k = \left\|\frac{1}{N}\sum_{i=1}^N g_k^i - \nabla F_i(w_{k-1})\right\|^2$, then expending the squared norm:

$$\begin{aligned}
\mathbb{E}\left[\square_k \mid \mathcal{H}_{k-1}\right] &= \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - \nabla F_i(w_{k-1})\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\
&\quad + \frac{1}{N^2}\sum_{i,j\in\{1,\dots,N\}/i\neq j} \underbrace{\mathbb{E}\left[\left\langle g_k^i - \nabla F_i(w_{k-1}), g_k^j - \nabla F_j(w_{k-1})\right\rangle \;\middle|\; \mathcal{H}_{k-1}\right]}_{=0 \text{ by independence of } (g_k^i)_{i=0}^N} \\
&= \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|(g_k^i - \nabla F_i(w_*)) + (\nabla F_i(w_*) - \nabla F_i(w_{k-1}))\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] .
\end{aligned}$$

Developing the squared norm a second time:

$$\begin{aligned}
\mathbb{E}\left[\square_k \mid \mathcal{H}_{k-1}\right] &= \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - \nabla F_i(w_*)\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\
&\quad + \frac{2}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\langle g_k^i - \nabla F_i(w_*), \nabla F_i(w_*) - \nabla F_i(w_{k-1})\right\rangle \;\middle|\; \mathcal{H}_{k-1}\right] \\
&\quad + \frac{1}{N^2}\sum_{i=1}^N \left\|\nabla F_i(w_{k-1}) - \nabla F_i(w_*)\right\|^2 \\
&= \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - \nabla F_i(w_*)\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] - \frac{1}{N^2}\sum_{i=1}^N \left\|\nabla F_i(w_{k-1}) - \nabla F_i(w_*)\right\|^2 \\
&\leq \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - \nabla F_i(w_*)\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] .
\end{aligned}$$

Recall that we note $h_*^i = \nabla F_i(w_*)$, returning to Equation (S3), we have:

$$\mathbb{E}\left[\left\|g_k\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] + L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle ,$$

which allows to conclude. ■

Appendix D.3. Lemmas for the case without memory

In this subsection, we give lemmas that are used only to demonstrate Theorem S4 (i.e. without memory). Lemma S9 is used to remove the uplink compression noise.

Lemma S9 (Expectation of the squared norm of the compressed gradient when no memory)**.** *In the case without memory, we have the following bound on the squared norm of the compressed gradient, for all k in $\mathbb{N}^*$:*

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{g}_k\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] &\leq \frac{\omega_{\text{up}}}{N^2}\sum_{i=0}^N \mathbb{E}\left[\left\|g_k^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] + \frac{1}{N^2}\sum_{i=0}^N \mathbb{E}\left[\left\|g_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\
&\quad + L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle .
\end{aligned}$$

Proof Let k in $\mathbb{N}^*$, first, we write as following:

$$\|\widehat{g}_k\|^2 = \|\widehat{g}_k - g_k\|^2 + 2\langle \widehat{g}_k - g_k, g_k\rangle + \|g_k\|^2 .$$

Taking stochastic expectation (recall that g_k is $\mathcal{F}_k$-measurable and that $\mathcal{H}_{k-1} \subset \mathcal{F}_k$):

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[\|\widehat{g}_k\|^2 \,\middle|\, \mathcal{F}_k\right] \,\middle|\, \mathcal{H}_{k-1}\right] &= \mathbb{E}\left[\mathbb{E}\left[\|\widehat{g}_k - g_k\|^2 \,\middle|\, \mathcal{F}_k\right] \,\middle|\, \mathcal{H}_{k-1}\right] \\ &\quad + 2 \times \mathbb{E}\left[\mathbb{E}\left[\langle \widehat{g}_k - g_k, g_k\rangle \mid \mathcal{F}_k\right] \mid \mathcal{H}_{k-1}\right] \\ &\quad + \mathbb{E}\left[\|g_k\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] .\end{aligned} \tag{S4}$$

We need to find a bound for each of the terms of above Equation (S4). The second term is zero in expectation and the last term is handled in Lemma S8. It follows that we just need to bound $\|\widehat{g}_k - g_k\|^2$:

$$\begin{aligned}\mathbb{E}\left[\|\widehat{g}_k - g_k\|^2 \,\middle|\, \mathcal{F}_k\right] &= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \widehat{g}_k^i - g_k^i\right\|^2 \,\middle|\, \mathcal{F}_k\right] \\ &= \frac{1}{N^2}\sum_{i=0}^N \mathbb{E}\left[\left\|\widehat{g}_k^i - g_k^i\right\|^2 \,\middle|\, \mathcal{F}_k\right] + \underbrace{\frac{1}{N}\sum_{i\neq j}\mathbb{E}\left[\left\langle \widehat{g}_k^i - g_k^i, \widehat{g}_k^j - g_k^j\right\rangle \,\middle|\, \mathcal{F}_k\right]}_{=0 \text{ because } (\widehat{g}_k^i)_{i=1}^N \text{ are independents}} \\ &= \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|\widehat{g}_k^i - g_k^i\right\|^2 \,\middle|\, \mathcal{F}_k\right] .\end{aligned}$$

Combining with Proposition S3, we hold that $\mathbb{E}\left[\|\widehat{g}_k - g_k\|^2 \,\middle|\, \mathcal{F}_k\right] \leq \frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^N \left\|g_k^i\right\|^2$. Furthermore, we have that:

- $\mathbb{E}\left[\langle \widehat{g}_k - g_k, g_k\rangle \mid \mathcal{F}_k\right] = 0$ (Proposition S2)
- $\mathbb{E}\left[\|g_k\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \leq \frac{1}{N^2}\sum_{i=0}^N \mathbb{E}\left[\left\|g_k^i - h_*^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] + L\langle \nabla F(w_{k-1}), w_{k-1} - w_*\rangle$ (Lemma S8).

Thus, we obtain from Equation (S4):

$$\begin{aligned}\mathbb{E}\left[\|\widehat{g}_k\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \leq \frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] + \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - h_*^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \\ + L\langle \nabla F(w_{k-1}), w_{k-1} - w_*\rangle .\end{aligned}$$

■

Lemma S10. *In the case without memory, we have the following bound on the squared norm of the local compressed gradient, for all k in $\mathbb{N}^*$, for all i in $[\![1, N]\!]$:* $\mathbb{E}[\|\widehat{g}_k^i\|^2 \mid \mathcal{F}_k] \leq (\omega_{\mathrm{up}} + 1)\|g_k^i\|^2$

Proof Let k in $\mathbb{N}^*$ and i in $[\![1, N]\!]$:

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{g}_k^i\right\|^2 \,\middle|\, \mathcal{F}_k\right] &= \mathbb{E}\left[\left\|\widehat{g}_k^i - g_k^i + g_k^i\right\|^2 \,\middle|\, \mathcal{F}_k\right] \\ &= \mathbb{E}\left[\left\|\widehat{g}_k^i - g_k^i\right\|^2 \,\middle|\, \mathcal{F}_k\right] + 2\underbrace{\mathbb{E}\left[\left\langle \widehat{g}_k^i - g_k^i, g_k^i\right\rangle \,\middle|\, \mathcal{F}_k\right]}_{=0} + \mathbb{E}\left[\left\|g_k^i\right\|^2 \,\middle|\, \mathcal{F}_k\right]\end{aligned}$$

We obtain the result because $\left\|g_k^i\right\|^2$ is $\mathcal{F}_{k+1}$-measurable and using Proposition S5. ■

Appendix D.4. Lemmas for the case with memory

In this Subsection, we give lemmas that are used only to demonstrate Theorems S5 and S6 (i.e. with memory). In order to derive an upper bound on the squared norm of $\|w_k - w_*\|^2$, for k in $\mathbb{N}^*$, we need to control $\|\widehat{g}_k\|^2$. This term is decomposed as a sum of three terms depending on:

1. the recursion over the memory term (h_{k-1}^i)

2. the difference between the stochastic gradient at the current point and at the optimal point (later controlled by co-coercivity)
3. the noise over stochasticity.

Lemma S11. *In the case with memory, we have the following upper bound on the squared norm of the compressed gradient, for all k in $\mathbb{N}^*$:*

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{g}_k\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq \frac{2(2\omega_{\mathrm{up}}+1)}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\ + \frac{2\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|h_{k-1}^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\ + L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle + \frac{2(2\omega_{\mathrm{up}}+1)\sigma_*}{Nb}\,.\end{aligned}$$

Proof

Let k in $\mathbb{N}^*$. We take the expectation w.r.t. the σ-algebra $\mathcal{H}_{k-1}$, with a bias-variance decomposition and we obtain $\mathbb{E}[\|\widehat{g}_k\|^2 \mid \mathcal{H}_{k-1}] = \mathbb{E}[\|g_k\|^2 \mid \mathcal{H}_{k-1}] + \mathbb{E}[\|\widehat{g}_k - g_k\|^2 \mid \mathcal{H}_{k-1}]$. The first term is handled with Lemma S8:

$$\mathbb{E}\left[\left\|g_k\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] + L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle\,.$$

Furthermore, by the independence of the "N" compressions:

$$\mathbb{E}\left[\left\|\widehat{g}_k - g_k\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] = \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\widehat{\Delta}_k^i - \Delta_k^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right]\,,$$

because $\mathcal{H}_{k-1} \subset \mathcal{F}_k$, we can use Proposition S3 to obtain $\mathbb{E}\left[\left\|\widehat{g}_k - g_k\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq \frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\left\|\Delta_k^i\right\|^2$ and next with Lemma S5, we have:

$$\mathbb{E}\left[\left\|\widehat{g}_k - g_k\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq \frac{2\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] + \mathbb{E}\left[\left\|h_{k-1}^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right]\,.$$

At the end:

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{g}_k\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] = \frac{2\omega_{\mathrm{up}}+1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] + \frac{2\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|h_{k-1}^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\ + L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle\,.\end{aligned}$$

We can now apply Lemma S7 to conclude the proof:

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{g}_k\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq \frac{2(2\omega_{\mathrm{up}}+1)}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\ + \frac{2\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|h_{k-1}^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\ + L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle + \frac{2(2\omega_{\mathrm{up}}+1)\sigma_*}{Nb}\,.\end{aligned}$$

■

To show that the Lyapunov function is a contraction, we need to find a bound for each terms. Bounding $\|w_k - w_*\|^2$, for k in $\mathbb{N}$, flows from update schema (see Equation (3)) decomposition. However the memory term $\left\|h_k^i - h_*^i\right\|^2$ involved in the Lyapunov function doesn't show up naturally. The aim of Lemma S12 is precisely to provide a recursive bound over the memory term to highlight the contraction. Like Lemma S6, the following lemma comes from Mishchenko et al. [29].

Lemma S12 (Recursive inequalities over memory term)**.** *Let $k \in \mathbb{N}^*$ and let $i \in [\![1, N]\!]$. The memory term used in the uplink broadcasting can be bounded using a recursion:*

$$\begin{aligned}\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq \left(1 + 2\alpha_{\mathrm{up}}^2\omega_{\mathrm{up}} + 2\alpha_{\mathrm{up}}^2 - 3\alpha_{\mathrm{up}}\right)\left\|h_{k-1}^i - h_*^i\right\|^2 \\ + 2(2\alpha_{\mathrm{up}}^2\omega_{\mathrm{up}} + 2\alpha_{\mathrm{up}}^2 - \alpha_{\mathrm{up}})\mathbb{E}\left[\left\|g_k - g_{k,*}\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\ + \frac{2\sigma_*^2}{b}\left(2\alpha_{\mathrm{up}}^2(\omega_{\mathrm{up}} + 1) - \alpha_{\mathrm{up}}\right) .\end{aligned}$$

Proof

Let $k \in \mathbb{N}^*$ and let $i \in [\![1, N]\!]$, using Lemma S3 we have:

$$\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{F}_k\right] = \left\|\mathbb{E}\left[h_k^i \mid \mathcal{F}_k\right] - h_*^i\right\|^2 + \mathbb{E}\left[\left\|h_k^i - \mathbb{E}\left[h_k^i \mid \mathcal{F}_k\right]\right\|^2 \;\middle|\; \mathcal{F}_k\right] ,$$

and now with Lemma S6:

$$\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{F}_k\right] = \left\|(1-\alpha_{\mathrm{up}})h_{k-1}^i + \alpha_{\mathrm{up}}g_k^i - h_*^i\right\|^2 + \mathbb{E}\left[\left\|h_k^i - \mathbb{E}\left[h_k^i \mid \mathcal{F}_k\right]\right\|^2 \;\middle|\; \mathcal{F}_k\right] .$$

Now recall that $h_k^i = h_{k-1}^i + \alpha_{\mathrm{up}}\widehat{\Delta}_k^i$, with $\mathbb{E}[\widehat{\Delta}_k^i \mid \mathcal{F}_k] = \Delta_k^i$ and h_{k-1}^i being $\mathcal{F}_k$-measurable:

$$\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{F}_k\right] = \left\|(1-\alpha_{\mathrm{up}})(h_{k-1}^i - h_*^i) + \alpha_{\mathrm{up}}(g_k^i - h_*^i)\right\|^2 + \alpha_{\mathrm{up}}^2\mathbb{E}\left[\left\|\widehat{\Delta}_k^i - \Delta_k^i\right\|^2 \;\middle|\; \mathcal{F}_k\right] .$$

Using Lemma S2 of Appendix D.1 and Proposition S3:

$$\begin{aligned}\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{F}_k\right] \leq (1-\alpha_{\mathrm{up}})\left\|h_{k-1}^i - h_*^i\right\|^2 + \alpha_{\mathrm{up}}\left\|g_k^i - h_*^i\right\|^2 \\ - \alpha_{\mathrm{up}}(1-\alpha_{\mathrm{up}})\left\|h_{k-1}^i - g_k^i\right\|^2 + \alpha_{\mathrm{up}}^2\omega_{\mathrm{up}}\left\|\Delta_k^i\right\|^2 .\end{aligned}$$

Because $h_{k-1}^i - g_k^i = \Delta_k^i$:

$$\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{F}_k\right] \leq (1-\alpha_{\mathrm{up}})\left\|h_{k-1}^i - h_*^i\right\|^2 + \alpha_{\mathrm{up}}\left\|g_k^i - h_*^i\right\|^2 + \alpha_{\mathrm{up}}\left(\alpha_{\mathrm{up}}(\omega_{\mathrm{up}}+1) - 1\right)\left\|\Delta_k^i\right\|^2 ,$$

and using Lemma S5:

$$\begin{aligned}\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{F}_k\right] &\leq (1-\alpha_{\mathrm{up}})\left\|h_{k-1}^i - h_*^i\right\|^2 + \alpha_{\mathrm{up}}\left\|g_k^i - h_*^i\right\|^2 \\ &\quad + 2\alpha_{\mathrm{up}}\left(\alpha_{\mathrm{up}}(\omega_{\mathrm{up}}+1) - 1\right)\left(\left\|h_{k-1}^i - h_*^i\right\|^2 + \left\|g_k - h_*^i\right\|^2\right) \\ &\leq \left(1 + 2\alpha_{\mathrm{up}}^2\omega_{\mathrm{up}} + 2\alpha_{\mathrm{up}}^2 - 3\alpha_{\mathrm{up}}\right)\left\|h_{k-1}^i - h_*^i\right\|^2 \\ &\quad + \alpha_{\mathrm{up}}(2\alpha_{\mathrm{up}}\omega_{\mathrm{up}} + 2\alpha_{\mathrm{up}} - 1)\left\|g_k - h_*^i\right\|^2 .\end{aligned}$$

Finally taking expectation w.r.t. the σ-algebra $\mathcal{H}_{k-1}$ ($\mathcal{H}_{k-1} \subset \mathcal{F}_k$) and using Equation (S2) of Lemma S7, we have:

$$\begin{aligned}\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq \left(1 + 2\alpha_{\mathrm{up}}^2\omega_{\mathrm{up}} + 2\alpha_{\mathrm{up}}^2 - 3\alpha_{\mathrm{up}}\right)\left\|h_{k-1}^i - h_*^i\right\|^2 \\ + 2(2\alpha_{\mathrm{up}}^2\omega_{\mathrm{up}} + 2\alpha_{\mathrm{up}}^2 - \alpha_{\mathrm{up}})\mathbb{E}\left[\left\|g_k - g_{k,*}\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\ + \frac{2\sigma_*^2}{b}\left(2\alpha_{\mathrm{up}}^2(\omega_{\mathrm{up}} + 1) - \alpha_{\mathrm{up}}\right) ,\end{aligned}$$

which concludes the proof.

■

After successfully invoking all previous lemmas, we will finally be able to use co-coercivity. Lemma S13 shows how Assumption 2 is used to do it. After this stage, proof will be continued by applying strong-convexity of F.

Lemma S13 (Applying co-coercivity)**.** *This lemma shows how to apply co-coercivity on stochastic gradients. For all k in $\mathbb{N}^*$, we have $\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - g_{k,*}\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle$.*

Proof Let $k \in \mathbb{N}^*$, using Assumption 2, we have:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \leq \frac{1}{N}\sum_{i=1}^{N} L\left\langle \mathbb{E}\left[g_k^i - g_{k,*}^i \mid \mathcal{H}_{k-1}\right], w_{k-1} - w_*\right\rangle$$
$$\leq L\left\langle \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(w_{k-1}) - \nabla F_i(w_*), w_{k-1} - w_*\right\rangle .$$

■

Appendix E. Proofs of Theorems

In this Section, we give demonstrations of all our theorems, that is to say, first the proofs of Theorems S4 and S5 from which flow Theorem 1. Their demonstration sketch is drawn from Mishchenko et al. [29]. And in a second time, we give a complete demonstration of Theorems 2 and 3. For the sake of demonstration, we define a Lyapunov function V_k [as in 29, 23], for k in $\mathbb{N}$:

$$V_k = \|w_k - w_*\|^2 + 2\gamma^2 C\frac{1}{N}\sum_{i=1}^{N}\left\|h_k^i - h_*^i\right\|^2 ,$$

with C in $\mathbb{R}_+^*$. The Lyapunov function is defined by combining two terms.

1. The distance from parameter w_k to optimal parameter w_*.
2. The memory term, the distance between the next element prediction h_k^i and the true gradient $h_*^i = \nabla F_i(w_*)$.

The aim is to proof that this function is a $(1-\gamma\mu)$ contraction for each variant of `Artemis`. To show that it's a contraction, we need three stages:

1. we develop the update schema defined in Equation (3) to get a first bound on $\|w_k - w_*\|^2$,
2. we find a recurrence over the memory term $\left\|h_k^i - h_*^i\right\|^2$,
3. and finally we combines the two equations to obtain the expected contraction using co-coercivity and strong-convexity.

Appendix E.1. Proof of main Theorem for `Artemis` - variant without memory

Theorem S4 (Unidirectional or bidirectional compression without memory)**.** *Considering that Assumptions 1 to 5 hold. Taking γ such that*

$$\gamma \leq \frac{N}{L(\omega_{\mathrm{dwn}}+1)\left(N + 2(\omega_{\mathrm{up}}+1)\right)} ,$$

then running `Artemis` with $\alpha_{\mathrm{up}} = 0$ (i.e without memory), we have for all k in $\mathbb{N}^*$:

$$\mathbb{E}\left\|w_k - w_*\right\|^2 \leq (1-\gamma\mu)^k\left\|w_0 - w_*\right\|^2 + 2\gamma\frac{E}{\mu N} ,$$

with $E = (\omega_{\mathrm{dwn}}+1)\left(\frac{(\omega_{\mathrm{up}}+1)\sigma_*^2}{b} + \omega_{\mathrm{up}}B^2\right)$. In the case of unidirectional compression (resp. no compression), we have $\omega_{\mathrm{dwn}} = 0$ (resp. $\omega_{\mathrm{up/dwn}} = 0$).

Proof

In the case of the variant of `Artemis` with $\alpha_{\mathrm{up}} = 0$, we don't have any memory term, thus $C = 0$ and we don't need to use the Lyapunov function.

Let k in $\mathbb{N}^*$, we start by writing that by definition of Equation (3):

$$\|w_k - w_*\|^2 = \|w_{k-1} - \gamma\Omega_k - w_*\|^2$$
$$= \|w_{k-1} - w_*\|^2 - 2\gamma\left\langle\Omega_k, w_{k-1} - w_*\right\rangle + \gamma^2\left\|\Omega_k\right\|^2 ,$$

with $\Omega_k = \mathcal{C}_{\mathrm{dwn}}(\widehat{g}_k)$ and $\widehat{g}_k = \frac{1}{N}\sum_{i=1}^{N}\widehat{g}_k^i$. First, we have $\mathbb{E}\left[\Omega_k \mid \mathcal{G}_{k-1}\right] = \widehat{g}_k$ (Proposition S4) secondly considering that $\mathbb{E}[\|\Omega_k\|^2 \mid \mathcal{G}_{k-1}] = \mathbb{V}(\Omega_k) + \|\mathbb{E}\left[\Omega_k \mid \mathcal{G}_{k-1}\right]\|^2 = (\omega_{\mathrm{dwn}}+1)\left\|\widehat{g}_k\right\|^2$ leads to:

$$\mathbb{E}\left[\left\|w_k - w_*\right\|^2 \,\middle|\, \mathcal{G}_{k-1}\right] = \mathbb{E}\left[\left\|w_{k-1} - w_*\right\|^2 \,\middle|\, \mathcal{G}_{k-1}\right] - 2\gamma\left\langle\widehat{g}_k, w_{k-1} - w_*\right\rangle + \gamma^2(\omega_{\mathrm{dwn}}+1)\left\|\widehat{g}_k\right\|^2 .$$

Now, we take expectation w.r.t σ-algebra $\mathcal{H}_{k-1} \subset \mathcal{G}_{k-1}$, (with use of Propositions S1 and S2, we obtain :

$$\mathbb{E}\left[\|w_k - w_*\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] = \mathbb{E}\left[\|w_{k-1} - w_*\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] - 2\gamma \left\langle \nabla F(w_{k-1}), w_{k-1} - w_* \right\rangle + \gamma^2(\omega_{\text{dwn}} + 1)\mathbb{E}\left[\|\widehat{g}_k\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] . \tag{S1}$$

Lemma S9 gives:

$$\mathbb{E}[\|\widehat{g}_k\|^2 \mid \mathcal{H}_{k-1}] \leq \frac{\omega_{\text{up}}}{N^2}\sum_{i=0}^{N}\mathbb{E}[\left\|g_k^i\right\|^2 \mid \mathcal{H}_{k-1}] + \frac{1}{N}\sum_{i=0}^{N}\mathbb{E}[\left\|g_k^i - h_*^i\right\|^2 \mid \mathcal{H}_{k-1}] + L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_* \right\rangle .$$

Lets introducing the noise at optimal point w_* with the two equations of Lemma S7:

$$\mathbb{E}\left[\|\widehat{g}_k\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq \frac{\omega_{\text{up}}}{N^2}\sum_{i=1}^{N} 2\left(\mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] + (\frac{\sigma_*^2}{b} + B^2)\right) + \frac{1}{N^2}\sum_{i=1}^{N} 2\left(\mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] + \frac{\sigma_*^2}{b}\right) + L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_* \right\rangle .$$

Invoking cocoercivity (Assumption 2):

$$\begin{aligned}\mathbb{E}\left[\|\widehat{g}_k\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] &\leq \frac{2(\omega_{\text{up}}+1)}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[L\left\langle g_k^i - g_{k,*}^i, w_{k-1} - w_*\right\rangle \mid \mathcal{H}_{k-1}\right] \\ &\quad + L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle + \frac{2}{N}\left(\frac{(\omega_{\text{up}}+1)\sigma_*^2}{b} + \omega_{\text{up}}B^2\right) \\ &\leq \frac{2(\omega_{\text{up}}+1)L}{N}\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle \\ &\quad + L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle + \frac{2}{N}\left(\frac{(\omega_{\text{up}}+1)\sigma_*^2}{b} + \omega_{\text{up}}B^2\right) .\end{aligned} \tag{S2}$$

Finally, we can inject Equation (S2) in Equation (S1) to obtain:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] &\leq \|w_{k-1} - w_*\|^2 \\ &\quad - 2\gamma\left(1 - \frac{\gamma L(\omega_{\text{dwn}}+1)(\omega_{\text{up}}+1)}{N} - \frac{\gamma L(\omega_{\text{dwn}}+1)}{2}\right)\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle \\ &\quad + \frac{2\gamma^2(\omega_{\text{dwn}}+1)\left(\frac{(\omega_{\text{up}}+1)\sigma_*^2}{b} + \omega_{\text{up}}B^2\right)}{N} .\end{aligned} \tag{S3}$$

We note:
1. $\square = 1 - \frac{\gamma L(\omega_{\text{dwn}}+1)(\omega_{\text{up}}+1)}{N} - \frac{\gamma L(\omega_{\text{dwn}}+1)}{2}$
2. $E = (\omega_{\text{dwn}}+1)\left(\frac{(\omega_{\text{up}}+1)\sigma_*^2}{b} + \omega_{\text{up}}B^2\right)$.

We need $\square \geq 0$ in order to further apply strong-convexity. However, in order to later obtain a convergence in $(1-\gamma\mu)$, we will use a stronger condition and, instead, state that we need $\square \geq 1/2$, which is equivalent to:

$$\frac{1}{2} \geq \frac{\gamma L(\omega_{\text{dwn}}+1)(\omega_{\text{up}}+1)}{N} + \frac{\gamma L(\omega_{\text{dwn}}+1)}{2} \iff \gamma \leq \frac{N}{L(\omega_{\text{dwn}}+1)\left(N + 2(\omega_{\text{up}}+1)\right)} ,$$

Using strong-convexity of F (Assumption 1), we rewrite Equation (S3) as follows:

$$\mathbb{E}\left[\|w_k - w_*\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq \|w_{k-1} - w_*\|^2 - 2\gamma\mu\square\|w_{k-1} - w_*\|^2 + 2\gamma^2\frac{E}{N} , \text{ equivalent to:}$$

$$\leq (1 - 2\gamma\mu\square)\|w_{k-1} - w_*\|^2 + 2\gamma^2\frac{E}{N} .$$

To guarantee a $(1-\gamma\mu)$ convergence, we need $\square \geq 1/2$, which is already verified, hence taking full expectation, we are allowed to write:

$$\begin{aligned} & \mathbb{E}[\|w_k - w_*\|^2] \leq (1-\gamma\mu)\mathbb{E}[\|w_{k-1} - w_*\|^2] + 2\gamma^2 \frac{E}{N} \\ \iff \quad & \mathbb{E}[\|w_k - w_*\|^2] \leq (1-\gamma\mu)^k \|w_0 - w_*\|^2 + 2\gamma^2 \frac{E}{N} \times \frac{1-(1-\gamma\mu)^k}{\gamma\mu} \\ \iff \quad & \mathbb{E}[\|w_k - w_*\|^2] \leq (1-\gamma\mu)^k \|w_0 - w_*\|^2 + 2\gamma \frac{E}{\mu N}\,, \end{aligned}$$

and the proof is complete.

■

Appendix E.2. Proof of main Theorem for `Artemis` *- variant with memory*

Theorem S5 (Unidirectional or bidirectional compression with memory)**.** *Considering that Assumptions 1 to 5 hold. We use w_* to indicate the optimal parameter such that $\nabla F(w_*) = 0$, and we note $h_*^i = \nabla F_i(w_*)$. We define the Lyapunov function for any k in $\mathbb{N}$:*

$$V_k = \|w_k - w_*\|^2 + 2\gamma^2 C \frac{1}{N} \sum_{i=1}^N \left\|h_k^i - h_*^i\right\|^2 .$$

We defined $C \in \mathbb{R}_+^*$, such that:

$$\frac{\omega_{\rm up}(\omega_{\rm dwn}+1)}{\alpha_{\rm up}\left(3 - 2\alpha_{\rm up}(\omega_{\rm up}+1)\right)} \leq C \leq \frac{N - \gamma L(\omega_{\rm dwn}+1)\left(N + 2(2\omega_{\rm up}+1)\right)}{4\gamma L \alpha_{\rm up}\left(2\alpha_{\rm up}(\omega_{\rm up}+1) - 1\right)} . \tag{S4}$$

Then, using `Artemis` *with a memory mechanism ($\alpha_{\rm up} \neq 0$), the convergence of the algorithm is guaranteed if:*

1. $\dfrac{1}{2(\omega_{\rm up}+1)} \leq \alpha_{\rm up} < \min\left(\dfrac{3}{2(\omega_{\rm up}+1)}, \quad \dfrac{3N - \gamma L(\omega_{\rm dwn}+1)\left(3N + 8\omega_{\rm up}+6\right)}{2(\omega_{\rm up}+1)(N - \gamma L(\omega_{\rm dwn}+1)(N+2))}\right)$
2. $$\gamma < \min\left\{ \begin{array}{c} \dfrac{1}{(\omega_{\rm dwn}+1)\left(1+\dfrac{2}{N}\right)L}, \quad \dfrac{3}{(\omega_{\rm dwn}+1)\left(3+\dfrac{8\omega_{\rm up}+6}{N}\right)L}, \\ \dfrac{N}{(\omega_{\rm dwn}+1)\left(N+2(2\omega_{\rm up}+1)\right)L} \end{array} \right\} .$$

And we have a bound for the Lyapunov function:

$$\mathbb{E}V_k \leq (1-\gamma\mu)^k \left(\|w_0 - w_*\|^2 + 2C\gamma^2 B^2\right) + 2\gamma \frac{E}{\mu N}\,,$$

with $E = \dfrac{\sigma_*^2}{b}\left((2\omega_{\rm up}+1)(\omega_{\rm dwn}+1) + 2C\left(2\alpha_{\rm up}^2(\omega_{\rm up}+1) - \alpha_{\rm up}\right)\right)$. In the case of unidirectional compression (resp. no compression), we have $\omega_{\rm dwn} = 0$ (resp. $\omega_{\rm up/dwn} = 0$).

Proof Let $k \in \mathbb{N}^*$, by definition of the update schema (Algorithm 1), we have: $w_k = w_{k-1} - \gamma\Omega_k$, with $\Omega_k = \mathcal{C}_{\rm dwn}(\widehat{g}_k)$ and $\widehat{g}_k = h_{k-1} + \frac{1}{N}\sum_{i=1}^N \widehat{\Delta}_k^i$, thus $\|w_k - w_*\|^2 = \|w_{k-1} - w_* + \gamma\Omega_k\|^2 = \|w_{k-1} - w_*\| - 2\gamma\left\langle \Omega_k, w_{k-1} - w_*\right\rangle + \gamma^2\|\Omega_k\|^2$. Taking expectation w.r.t. the σ-algebra $\mathcal{G}_{k-1}$:

$$\mathbb{E}\left[\|w_k - w_*\|^2 \;\middle|\; \mathcal{G}_{k-1}\right] = \mathbb{E}\left[\|w_{k-1} - w_*\|^2 \;\middle|\; \mathcal{G}_{k-1}\right] - 2\gamma\left\langle \widehat{g}_k, w_{k-1} - w_*\right\rangle + \gamma^2(\omega_{\rm dwn}+1)\|\widehat{g}_k\|^2 .$$

We take expectation w.r.t σ-algebra $\mathcal{H}_{k-1} \subset \mathcal{G}_{k-1}$ and invoke Lemma S11:

$$\begin{aligned}\mathbb{E}\left[\left\|w_k - w_*\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq \left\|w_{k-1} - w_*\right\|^2 &- 2\gamma\mathbb{E}\left[\langle \widehat{g}_k, w_{k-1} - w_*\rangle \mid \mathcal{H}_{k-1}\right] \\ &+ \frac{2(2\omega_{\rm up}+1)(\omega_{\rm dwn}+1)\gamma^2}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\ &+ \frac{2\omega_{\rm up}(\omega_{\rm dwn}+1)\gamma^2}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|h_{k-1}^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\ &+ \gamma^2(\omega_{\rm dwn}+1)L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle \\ &+ \frac{2(2\omega_{\rm up}+1)(\omega_{\rm dwn}+1)\gamma^2\sigma_*}{Nb}. \end{aligned} \tag{S5}$$

Note that in the case of unidirectional compression, we have $\Omega_k = \widehat{g}_k$, and the steps above are more straightforward. Recall that according to Lemma S12 (and taking the sum), we have:

$$\begin{aligned}&\frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\ &\quad\leq \left(1 + 2\alpha_{\rm up}^2\omega_{\rm up} + 2\alpha_{\rm up}^2 - 3\alpha_{\rm up}\right)\frac{1}{N^2}\sum_{i=1}^N \left\|h_{k-1}^i - h_*^i\right\|^2 \\ &\quad+ 2(2\alpha_{\rm up}^2\omega_{\rm up} + 2\alpha_{\rm up}^2 - \alpha_{\rm up})\frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\ &\quad+ \frac{2\sigma_*^2}{Nb}\left(2\alpha_{\rm up}^2(\omega_{\rm up}+1) - \alpha_{\rm up}\right). \end{aligned} \tag{S6}$$

With a linear combination (S5) + $2\gamma^2 C$ (S6):

$$\begin{aligned}&\mathbb{E}\left[\left\|w_k - w_*\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] + 2\gamma^2 C\frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \\ &\quad\leq \left\|w_{k-1} - w_*\right\|^2 - 2\gamma\mathbb{E}\left[\langle \widehat{g}_k, w_{k-1} - w_*\rangle \mid \mathcal{H}_{k-1}\right] \\ &\qquad+ 2\gamma^2\frac{(2\omega_{\rm up}+1)(\omega_{\rm dwn}+1) + 2C(2\alpha_{\rm up}^2\omega_{\rm up} + 2\alpha_{\rm up}^2 - \alpha_{\rm up})}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i - g_{k,*}^i\right\| \mid \mathcal{H}_{k-1}\right] \\ &\qquad+ 2\gamma^2 C\left(\frac{\omega_{\rm up}(\omega_{\rm dwn}+1)}{C} + 1 + 2\alpha_{\rm up}^2\omega_{\rm up} + 2\alpha_{\rm up}^2 - 3\alpha_{\rm up}\right)\times\frac{1}{N^2}\sum_{i=1}^N \left\|h_{k-1}^i - h_*^i\right\|^2 \\ &\qquad+ \gamma^2(\omega_{\rm dwn}+1)L\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle \\ &\qquad+ \frac{2\gamma^2}{N}\left(\frac{\sigma_*^2}{b}\left((2\omega_{\rm up}+1)(\omega_{\rm dwn}+1) + 2C\left(2\alpha_{\rm up}^2(\omega_{\rm up}+1) - \alpha_{\rm up}\right)\right)\right). \end{aligned}$$

We transform $\|g_k^i - g_{k,*}^i\|^2$ applying co-coercivity (Lemma S13) and note:

- $\square = 1 - \gamma L(\omega_{\rm dwn}+1)/2 - \gamma L\left((2\omega_{\rm up}+1)(\omega_{\rm dwn}+1) + 2C(2\alpha_{\rm up}^2\omega_{\rm up} + 2\alpha_{\rm up}^2 - \alpha_{\rm up})\right)/N$
- $\diamondsuit = \dfrac{\omega_{\rm up}(\omega_{\rm dwn}+1)}{C} + 1 + 2\alpha_{\rm up}^2\omega_{\rm up} + 2\alpha_{\rm up}^2 - 3\alpha_{\rm up}$
- $E = \dfrac{\sigma_*^2}{b}\left((2\omega_{\rm up}+1)(\omega_{\rm dwn}+1) + 2C\left(2\alpha_{\rm up}^2(\omega_{\rm up}+1) - \alpha_{\rm up}\right)\right)$.

Now, because $\mathbb{E}\left[\widehat{g}_k \mid \mathcal{H}_{k-1}\right] = \mathbb{E}\left[h_{k-1} + \frac{1}{N}\sum_{i=1}^N \widehat{\Delta}_k^i \;\middle|\; \mathcal{H}_{k-1}\right] = \nabla F(w_{k-1})$, we have:

$$\begin{aligned}\mathbb{E}\left[V_k \mid \mathcal{H}_{k-1}\right] \leq \left\|w_{k-1} - w_*\right\|^2 &- 2\gamma\square\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle \\ &+ \frac{2\gamma^2\diamondsuit}{N^2}\sum_{i=1}^N \left\|h_{k-1}^i - h_*^i\right\|^2 + \frac{2\gamma^2 E}{N}. \end{aligned} \tag{S7}$$

Now, the goal is to apply strong-convexity of F (Assumption 1) using the inequality presented in Lemma S4. But then we must have $\square \geq 0$. However, in order to later obtain a convergence in $(1-\gamma\mu)$, we will use a stronger condition and, instead, state that we need $\square \geq 1/2$, which is equivalent to:

$$\begin{aligned}
&\frac{\omega_{\text{dwn}}+1}{2} + \frac{1}{N}\left((2\omega_{\text{up}}+1)(\omega_{\text{dwn}}+1) + 2C(2\alpha_{\text{up}}^2\omega_{\text{up}} + 2\alpha_{\text{up}}^2 - \alpha_{\text{up}})\right) \leq \frac{1}{2\gamma L} \\
\iff \quad & (2\omega_{\text{up}}+1)(\omega_{\text{dwn}}+1) + 2C(2\alpha_{\text{up}}^2\omega_{\text{up}} + 2\alpha_{\text{up}}^2 - \alpha_{\text{up}}) \leq \frac{(1-\gamma L(\omega_{\text{dwn}}+1))N}{2\gamma L} \\
\iff \quad & C \leq \frac{N - \gamma L(\omega_{\text{dwn}}+1)\left(N + 2(2\omega_{\text{up}}+1)\right)}{4\gamma L\alpha_{\text{up}}\left(2\alpha_{\text{up}}(\omega_{\text{up}}+1)-1\right)} .
\end{aligned}$$

This holds only if the numerator and the denominator are positive:

$$\begin{cases}
N - \gamma L(\omega_{\text{dwn}}+1)\left(N + 2(2\omega_{\text{up}}+1)\right) > 0 \iff \gamma < \dfrac{N}{(\omega_{\text{dwn}}+1)\left(N+2(2\omega_{\text{up}}+1)\right)L} \\
2\alpha_{\text{up}}(\omega_{\text{up}}+1) - 1 \leq 0 \iff \alpha_{\text{up}} \geq \dfrac{1}{2(\omega_{\text{up}}+1)} .
\end{cases}$$

strong-convexity is applied, and we obtain:

$$\mathbb{E}\left[V_k \mid \mathcal{H}_{k-1}\right] \leq (1 - 2\gamma\mu\square)\left\|w_{k-1} - w_*\right\|^2 + \frac{2\gamma^2 C\diamondsuit}{N}\sum_{i=1}^{N}\left\|h_{k-1}^i - h_*^i\right\|^2 + \frac{2\gamma^2 E}{N} . \tag{S8}$$

To guarantee a $(1-\gamma\mu)$ convergence, constants must verify: (1) $\square \geq 1/2$ and (2) $\diamondsuit \leq 1-\gamma\mu$. The first condition is already verified, and the second one leads to:

$$\begin{aligned}
\diamondsuit \leq 1-\gamma\mu &\iff \frac{\omega_{\text{dwn}}+1}{C}\omega_{\text{up}} \leq 3\alpha_{\text{up}} - 2\alpha_{\text{up}}^2\omega_{\text{up}} - 2\alpha_{\text{up}} - \gamma\mu \\
&\iff C \geq \frac{\omega_{\text{up}}(\omega_{\text{dwn}}+1)}{\alpha_{\text{up}}\left(3 - 2\alpha_{\text{up}}(\omega_{\text{up}}+1)\right) - \gamma\mu} .
\end{aligned}$$

In the following we will consider that $\frac{\gamma\mu}{\alpha_{\text{up}}} = \underset{\mu\to 0}{o}(1)$ which is possible because α_{up} is independent of μ (it depends only of ω_{up} and ω_{dwn}) and it result to:

$$\alpha_{\text{up}}\left(3 - 2\alpha_{\text{up}}(\omega_{\text{up}}+1)\right) - \gamma\mu \underset{\mu\to 0}{\sim} \alpha_{\text{up}}\left(3 - 2\alpha_{\text{up}}(\omega_{\text{up}}+1)\right)$$

Thus, the condition on C becomes $\dfrac{\omega_{\text{up}}(\omega_{\text{dwn}}+1)}{\alpha_{\text{up}}\left(3 - 2\alpha_{\text{up}}(\omega_{\text{up}}+1)\right)} \leq C$, which is correct only if $\alpha_{\text{up}} \leq \dfrac{3}{2(\omega_{\text{up}}+1)}$. And we obtain the following conditions on C:

$$\frac{\omega_{\text{up}}(\omega_{\text{dwn}}+1)}{\alpha_{\text{up}}\left(3 - 2\alpha_{\text{up}}(\omega_{\text{up}}+1)\right)} \leq C \leq \frac{N - \gamma L(\omega_{\text{dwn}}+1)\left(N + 2(2\omega_{\text{up}}+1)\right)}{4\gamma L\alpha_{\text{up}}\left(2\alpha_{\text{up}}(\omega_{\text{up}}+1)-1\right)} .$$

It follows, that the above interval is not empty if:

$$\frac{\omega_{\text{up}}(\omega_{\text{dwn}}+1)}{\alpha_{\text{up}}\left(3 - 2\alpha_{\text{up}}(\omega_{\text{up}}+1)\right)} \leq \frac{N - \gamma L(\omega_{\text{dwn}}+1)\left(N + 2(2\omega_{\text{up}}+1)\right)}{4\gamma L\alpha_{\text{up}}\left(2\alpha_{\text{up}}(\omega_{\text{up}}+1)-1\right)} .$$

For sake of clarity we denote momentarily $\widetilde{\gamma} = (\omega_{\text{dwn}}+1)\gamma L$, hence the above condition becomes:

$$\begin{aligned}
&8\alpha_{\text{up}}\omega_{\text{up}}(\omega_{\text{up}}+1)\widetilde{\gamma} - 4\omega_{\text{up}}\widetilde{\gamma} \leq 3N - 3\widetilde{\gamma}\left(N + 2 + 2(2\omega_{\text{up}}+1)\right)) \\
&\qquad\qquad - 2\alpha_{\text{up}}(\omega_{\text{up}}+1)N + 2\alpha_{\text{up}}\widetilde{\gamma}(\omega_{\text{up}}+1)\left(N + 2(2\omega_{\text{up}}+1)\right) \\
\iff \quad & 2\alpha_{\text{up}}(\omega_{\text{up}}+1)(N - \widetilde{\gamma}(N+2)) \leq 3N - \widetilde{\gamma}\left(3N + 8\omega_{\text{up}} + 6\right) .
\end{aligned}$$

And at the end, we obtain:

$$\alpha_{\text{up}} \leq \frac{3N - \gamma L(\omega_{\text{dwn}}+1)\left(3N + 8(\omega_{\text{up}}+6)\right)}{2(\omega_{\text{up}}+1)(N - \gamma L(\omega_{\text{dwn}}+1)(N+2))} .$$

Again, this implies two conditions on γ:

$$\begin{cases} 3N - \gamma L(\omega_{\text{dwn}}+1)\left(3N+8\omega_{\text{up}}+6\right) > 0 \iff \gamma < \dfrac{3}{(\omega_{\text{dwn}}+1)\left(3+\dfrac{8\omega_{\text{up}}+6}{N}\right)L} \\ N - \gamma L(\omega_{\text{dwn}}+1)(N+2) > 0 \iff \gamma < \dfrac{1}{(\omega_{\text{dwn}}+1)\left(1+\dfrac{2}{N}\right)L}\,. \end{cases}$$

The constant C exists, and from Equation (S8), taking full expectation, we are allowed to write $\mathbb{E}[V_k] \leq (1-\gamma\mu)\mathbb{E}[V_{k-1}] + 2\gamma^2\frac{E}{N}$. Unrolling the inequality we obtain:

$$\begin{aligned} & \mathbb{E}[V_k] \leq (1-\gamma\mu)^k \mathbb{E}V_0 + 2\gamma^2\frac{E}{N} \times \frac{1-(1-\gamma\mu)^k}{\gamma\mu} \\ \implies \quad & \mathbb{E}[V_k] \leq (1-\gamma\mu)^k V_0 + 2\gamma\frac{E}{\mu N}\,. \end{aligned}$$

Because $V_0 = \mathbb{E}\left\|w_0 - w_*\right\|^2 + 2\gamma^2 C\frac{1}{N}\sum_{i=0}^N \left\|h_*^i\right\|^2 \leq \left\|w_0 - w_*\right\|^2 + 2C\gamma^2B^2$ (Assumption 4), we can write:

$$\mathbb{E}[V_k] = (1-\gamma\mu)^k\left(\left\|w_0 - w_*\right\|^2 + 2C\gamma^2B^2\right) + 2\gamma\frac{E}{\mu N}\,.$$

Thus, we highlighted that the Lyapunov function V_k is a $(1-\gamma\mu)$ contraction if C is taken in a given interval, with γ and α_{up} satisfying some conditions. This guarantees the convergence of the `Artemis` using version 1 or 2 with $\alpha_{\text{up}} \neq 0$ (algorithm with uni-compression or bi-compression combined with a memory mechanism).

■

Appendix E.3. Proof of Theorem 2 - Polyak-Ruppert averaging

Theorem S6 (Unidirectional or bidirectional compression using memory and averaging)**.** *Considering now that F is convex, thus $\mu = 0$ and considering that Assumptions 2 to 5 hold. We use w_* to indicate the optimal parameter such that $\nabla F(w_*) = 0$, and we note $h_*^i = \nabla F_i(w_*)$. A Lyapunov function is defined for any k in $\mathbb{N}$:*

$$V_k = \left\|w_k - w_*\right\|^2 + 2\gamma^2 C\frac{1}{N}\sum_{i=1}^N \left\|h_k^i - h_*^i\right\|^2\,.$$

We defined $C \in \mathbb{R}_+^*$, such that:

$$\frac{\omega_{\text{up}}(\omega_{\text{dwn}}+1)}{\alpha_{\text{up}}\left(3-2\alpha_{\text{up}}(\omega_{\text{up}}+1)\right)} \leq C \leq \frac{N-\gamma L(\omega_{\text{dwn}}+1)\left(N+2(2\omega_{\text{up}}+1)\right)}{4\gamma L\alpha_{\text{up}}\left(2\alpha_{\text{up}}(\omega_{\text{up}}+1)-1\right)}\,.$$

Then running the variant of `Artemis` with $\alpha_{\text{up}} \neq 0$, hence with a memory mechanism, and using Polyak-Ruppert averaging, the convergence of the algorithm is guaranteed if:

1. $\dfrac{1}{2(\omega_{\text{up}}+1)} \leq \alpha_{\text{up}} < \min\left(\dfrac{3}{2(\omega_{\text{up}}+1)},\quad \dfrac{3N-\gamma L(\omega_{\text{dwn}}+1)\left(3N+8\omega_{\text{up}}+6\right)}{2(\omega_{\text{up}}+1)(N-\gamma L(\omega_{\text{dwn}}+1)(N+2))}\right)$
2. $$\gamma < \min\left\{\begin{array}{l} \dfrac{1}{(\omega_{\text{dwn}}+1)\left(1+\dfrac{2}{N}\right)L}, \quad \dfrac{3}{(\omega_{\text{dwn}}+1)\left(3+\dfrac{8\omega_{\text{up}}+6}{N}\right)L}, \\ \dfrac{N}{(\omega_{\text{dwn}}+1)\left(N+2(2\omega_{\text{up}}+1)\right)L} \end{array}\right\}\,. \tag{S9}$$

And we have the following bound for the Polyak-Ruppert averaged iterate $\overline{w}_{K-1} = \frac{1}{K}\sum_{k=0}^{K-1} w_k$*:*

$$\mathbb{E}\left[F(\overline{w}_{K-1}) - F(w_*)\right] \leq \frac{\|w_0 - w_*\|^2 + 2C\gamma^2 B^2}{\gamma K} + 2\gamma\frac{E}{N}\,, \tag{S10}$$

with $E = \frac{\sigma_*^2}{b}\left((2\omega_{\text{up}}+1)(\omega_{\text{dwn}}+1) + 2C\left(2\alpha_{\text{up}}^2(\omega_{\text{up}}+1) - \alpha_{\text{up}}\right)\right)$*. Equation (S10) can be written as in Theorem 2 if we take* $\gamma = \min\left(\sqrt{\frac{N\delta_0^2}{2EK}}; \gamma_{\max}\right)$*, where* $\gamma_{\max}$ *is the maximal possible value of* γ *as precised by Equation (S9):*

$$\mathbb{E}\left[F(\overline{w}_{K-1}) - F(w_*)\right] \leq 2\max\left(\sqrt{\frac{2\delta_0^2 E}{NK}}; \frac{\delta_0^2}{\gamma_{\max}K}\right) + \frac{2\gamma_{\max}CB^2}{K}$$

Proof

Let k in $\mathbb{N}^*$, starting from Equation (S7) from the proof of Theorem S5, we have:

$$\mathbb{E}\left[V_k \mid \mathcal{H}_{k-1}\right] \leq \|w_{k-1} - w_*\|^2 - 2\gamma\square\left\langle \nabla F(w_{k-1}), w_{k-1} - w_*\right\rangle + \frac{2\gamma^2\diamondsuit}{N^2}\sum_{i=1}^{N}\left\|h_{k-1}^i - h_*^i\right\|^2 + \frac{2\gamma^2 E}{N}\,.$$

But this time, instead of applying strong-convexity of F, we apply convexity (Assumption 1 but with $\mu = 0$):

$$\mathbb{E}V_k \leq \|w_{k-1} - w_*\|^2 - 2\gamma\square\left(F(w_{k-1}) - F(w_*)\right) + \frac{2\gamma^2 C\diamondsuit}{N^2}\sum_{i=1}^{N}\left\|h_{k-1}^i - h_*^i\right\|^2 + \frac{2\gamma^2}{N}E \tag{S11}$$

As in Theorem S5, we want $\square \geq 1/2$, which is equivalent to:

$$\begin{aligned} &\frac{\omega_{\text{dwn}}+1}{2} + \frac{1}{N}\left((2\omega_{\text{up}}+1)(\omega_{\text{dwn}}+1) + 2C(2\alpha_{\text{up}}^2\omega_{\text{up}} + 2\alpha_{\text{up}}^2 - \alpha_{\text{up}})\right) \leq \frac{1}{2\gamma L} \\ \iff \quad & C \leq \frac{N - \gamma L(\omega_{\text{dwn}}+1)\left(N + 8\omega_{\text{up}} + 6\right)}{4\gamma L\alpha_{\text{up}}\left(2\alpha_{\text{up}}(\omega_{\text{up}}+1) - 1\right)}\,. \end{aligned} \tag{S12}$$

It holds only if the numerator and the denominator are positive:

$$\begin{cases} N - \gamma L(\omega_{\text{dwn}}+1)\left(N + 8\omega_{\text{up}} + 6\right) > 0 \iff \gamma < \dfrac{N}{(\omega_{\text{dwn}}+1)\left(N + 8\omega_{\text{up}} + 6\right)L} \\ 2\alpha_{\text{up}}(\omega_{\text{up}}+1) - 1 \leq 0 \iff \alpha_{\text{up}} \geq \dfrac{1}{2(\omega_{\text{up}}+1)}\,. \end{cases}$$

Returning to Equation (S11), taking benefit of Equation (S12) and passing $F(w_{k-1}) - F(w_*)$ on the left side gives:

$$\gamma(F(w_{k-1}) - F(w_*)) \leq \|w_{k-1} - w_*\|^2 + \frac{2\gamma^2 C\diamondsuit}{N^2}\sum_{i=1}^{N}\left\|h_{k-1}^i - h_*^i\right\|^2 - \mathbb{E}V_k + \frac{2\gamma^2}{N}E\,.$$

If $\diamondsuit \leq 1$, we have $\gamma\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] \leq \mathbb{E}V_{k-1} - \mathbb{E}V_k + 2\gamma^2 E/N$, and summing over all K in $\mathbb{N}^*$ iterations gives:

$$\begin{aligned} \gamma\left(\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right]\right) &\leq \frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{E}V_{k-1} - \mathbb{E}V_k + 2\gamma^2\frac{E}{N}\right) \\ &\leq \frac{\mathbb{E}V_0 - \mathbb{E}V_K}{K} + 2\gamma^2\frac{E}{N} \quad \text{because } E \text{ is independent of } K. \end{aligned}$$

Thus, by convexity:

$$\mathbb{E}\left[F\left(\frac{1}{K}\sum_{k=1}^{K}w_{k-1}\right) - F(w_*)\right] \leq \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] \leq \frac{V_0}{\gamma K} + 2\gamma\frac{E}{N}\,.$$

Last step is to extract conditions over γ and α_{up} from requirement $\diamondsuit \leq 1$:

$$\diamondsuit < 1 \iff \frac{2\omega_{\text{up}}(\omega_{\text{dwn}}+1)}{2C} < 3\alpha_{\text{up}} - 2\alpha_{\text{up}}^2\omega_{\text{up}} - 2\alpha_{\text{up}} \iff C > \frac{\omega_{\text{up}}(\omega_{\text{dwn}}+1)}{\alpha_{\text{up}}\left(3 - 2\alpha_{\text{up}}(\omega_{\text{up}}+1)\right)}\,,$$

and the second inequality is correct only if $\alpha_{\rm up} \leq \frac{3}{2(\omega_{\rm up}+1)}$. From this development follows the following conditions on C, which are equivalent to those obtain in Theorem S5

$$\frac{\omega_{\rm up}(\omega_{\rm dwn}+1)}{\alpha_{\rm up}\left(3-2\alpha_{\rm up}(\omega_{\rm up}+1)\right)} \leq C \leq \frac{N-\gamma L(\omega_{\rm dwn}+1)\left(N+2(2\omega_{\rm up}+1)\right)}{4\gamma L\alpha_{\rm up}\left(2\alpha_{\rm up}(\omega_{\rm up}+1)-1\right)}\,.$$

This interval is not empty:

$$\begin{aligned} &\frac{\omega_{\rm up}(\omega_{\rm dwn}+1)}{\alpha_{\rm up}\left(3-2\alpha_{\rm up}(\omega_{\rm up}+1)\right)} \leq \frac{N-\gamma L(\omega_{\rm dwn}+1)\left(N+2(2\omega_{\rm up}+1)\right)}{4\gamma L\alpha_{\rm up}\left(2\alpha_{\rm up}(\omega_{\rm up}+1)-1\right)} \\ \iff \quad &\alpha_{\rm up} \leq \frac{3N-\gamma L(\omega_{\rm dwn}+1)\left(3N+8\omega_{\rm up}+6\right)}{2(\omega_{\rm up}+1)(N-\gamma L(\omega_{\rm dwn}+1)(N+2))}\,. \end{aligned}$$

Again, this implies two conditions on γ:

$$\begin{cases} 3N-\gamma L(\omega_{\rm dwn}+1)\left(3N+8\omega_{\rm up}+6\right)>0 \iff \gamma < \dfrac{3}{(\omega_{\rm dwn}+1)\left(3+\dfrac{8\omega_{\rm up}+6}{N}\right)L} \\ N-\gamma L(\omega_{\rm dwn}+1)(N+2)>0 \iff \gamma < \dfrac{1}{(\omega_{\rm dwn}+1)\left(1+\dfrac{2}{N}\right)L}\,. \end{cases}$$

which guarantees the existence of C and thus the validity of the above development. In conclusion:

$$\begin{aligned} \mathbb{E}\left[F(\overline{w}_{K-1})-F(w_*)\right] &\leq \frac{V_0}{\gamma K}+2\gamma\frac{E}{N} \leq \frac{\|w_0-w_*\|^2+2C\gamma^2B^2}{\gamma K}+2\gamma\frac{E}{N} \\ &\leq \frac{\|w_0-w_*\|^2}{\gamma K}+2\gamma\left(\frac{E}{N}+\frac{CB^2}{K}\right)\,. \end{aligned}$$

Next, our goal is to define the optimal step-size γ_{opt}. With this aim, we bound $2\gamma\frac{CB^2}{K}$ by $2\gamma_{\max}\frac{CB^2}{K}$. This leads to ignore this term when optimizing the step-size and thus to obtain a simpler expression of γ_{opt}. This approximation is relevant, because B^2/K is "small". And we obtain:

$$\mathbb{E}\left[F(\overline{w}_{K-1})-F(w_*)\right] \leq \frac{\|w_0-w_*\|^2}{\gamma K}+2\gamma\frac{E}{N}+2\gamma_{\max}\frac{CB^2}{K}.$$

This is valid for all variants of `Artemis`, with step-size in Table 3 and E in Theorem 1. Subsequently, the "optimal" step-size (at least the one minimizing the upper bound) is

$$\gamma_{\rm opt}=\sqrt{\frac{\|w_0-w_*\|^2 N}{2EK}}\,,$$

resulting in a convergence rate as $2\sqrt{\frac{2\|w_0-w_*\|^2E}{NK}}+\frac{2\gamma_{\max}CB^2}{K}$, if this step-size is allowed. If $\sqrt{\frac{\|w_0-w_*\|^2N}{2EK}} \geq \gamma_{\max}$ $\left(\iff \frac{2\gamma_{\max}E}{N} \leq \frac{\|w_0-w_*\|^2}{\gamma_{\max}K}\right)$, then the bias term dominates and the upper bound is $2\frac{\|w_0-w_*\|^2}{\gamma_{\max}K}+\frac{2\gamma_{\max}CB^2}{K}$. Overall, the convergence rate is given by:

$$\mathbb{E}\left[F(\overline{w}_{K-1})-F(w_*)\right] \leq 2\max\left(\sqrt{\frac{2\|w_0-w_*\|^2E}{NK}};\frac{\|w_0-w_*\|^2}{\gamma_{\max}K}\right)+\frac{2\gamma_{\max}CB^2}{K}\,.$$

■

Appendix E.4. Proof of Theorem 3 - convergence in distribution

In this Section, we give the proof of Theorem 3. The theorem is decomposed into two main points, that are respectively derived from Propositions S7 and S8, given in Appendices E.4.2 and E.4.3. Throughout this Section, we consider a *linear* compression operator $\mathcal{C}$, for instance sparsification, then for any $z,z'\in\mathbb{R}^d$, we have that $\mathcal{C}(z)-\mathcal{C}(z')=\mathcal{C}(z-z')$. We first introduce a few notations in Appendix E.4.1.

Appendix E.4.1. Background on distributions and Markov chains

We consider `Artemis` iterates $(w_{k-1}, (h^i_{k-1})_{i\in[\![1,N]\!]})_{k\in\mathbb{N}} \in \mathbb{R}^{d(1+N)}$ with the following update equation:

$$\left\{\begin{array}{lll} & w_k & = w_{k-1} - \gamma \mathcal{C}_{\text{dwn}}\left(\frac{1}{N}\sum_{i=1}^N \mathcal{C}_{\text{up}}\left(g^i_k - h^i_{k-1}\right) + h^i_{k-1}\right) \\ \forall i \in [\![1,N]\!], & h^i_k & = h^i_{k-1} + \alpha_{\text{up}}\mathcal{C}_{\text{up}}\left(g^i_k - h^i_{k-1}\right) \end{array}\right. \tag{S13}$$

We see the iterates, for a constant step-size γ, as a homogeneous Markov chain, and denote $R_{\gamma,v}$ the *Markov kernel*, which is the equivalent for continuous spaces of the *transition matrix* in finite state spaces. Let $R_{\gamma,v}$ be the Markov kernel on $(\mathbb{R}^{d(1+N)}, \mathcal{B}(\mathbb{R}^{d(1+N)}))$ associated with the SGD iterates $(w_{k-1}, \tau(h^i_{k-1})_{i\in[\![1,N]\!]})_{k\geq 0}$ for a variant v of `Artemis`, as defined in Algorithm 1 and with τ a constant specified afterwards, where $\mathcal{B}(\mathbb{R}^{d(1+N)})$ is the Borel σ-field of $\mathbb{R}^{d(1+N)}$. Meyn & Tweedie [28] provide an introduction to Markov chain theory. For readability, we now denote $(h^i_{k-1})_i$ for $(h^i_{k-1})_{i\in[\![1,N]\!]}$.

Definition 3. *For any initial distribution ν_0 on $\mathcal{B}(\mathbb{R}^{d(1+N)})$ and $k \in \mathbb{N}^*$, $\nu_0 R^k_{\gamma,v}$ denotes the distribution of $(w_{k-1}, \tau(h^i_{k-1})_i)$ starting at $(w_0, \tau(h^i_0)_i)$ distributed according to ν_0.*

We can make the following comments:

1. **Initial distribution.** We consider deterministic initial points, i.e., $(w_0, \tau(h^i_0)_i)$ follows a Dirac at point $(w_0, \tau(h^i_0)_i)$. We denote this Dirac $\delta_{w_0} \otimes \otimes_{i=1}^N \delta_{\tau h^i_0} \overset{\text{not.}}{=} \delta_{w_0} \otimes \delta_{\tau h^1_0} \otimes \cdots \otimes \delta_{\tau h^N_0}$.
2. **Notation in the main text:** In the main text, for simplicity, we used Θ_k to denote the distribution of w_{k-1} when launched from $(w_0, \tau(h^i_0)_i)$. Thus Θ_k corresponds to the distribution of the projection on first d coordinates of $((\delta_{w_0} \otimes \otimes_{i=1}^N \delta_{\tau h^i_0}) R^k_\gamma)$.
3. **Case without memory:** In the memory-less case, we have $(h^i_{k-1})_{k\in\mathbb{N}} \equiv 0$, and could restrict ourselves to a Markov kernel on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

For any variant v of `Artemis`, we prove that $(w_{k-1}, (h^i_{k-1})_i)_{k\geq 0}$ admits a limit stationary distribution

$$\Pi_{\gamma,v} = \pi_{\gamma,v,w} \otimes \pi_{\gamma,v,(h)} \tag{S14}$$

and quantify the convergence of $((\delta_{w_0} \otimes \otimes_{i=1}^N \delta_{\tau h^i_0}) R^k_\gamma)_{k\geq 0}$ to $\Pi_{\gamma,v}$, in terms of Wasserstein metric $\mathcal{W}_2$.

Definition 4. *For all probability measures ν and λ on $\mathcal{B}(\mathbb{R}^d)$, such that $\int_{\mathbb{R}^d} \|w\|^2 \,\mathrm{d}\nu(w) < +\infty$ and $\int_{\mathbb{R}^d} \|w\|^2 \,\mathrm{d}\lambda(w) \leq +\infty$, define the squared Wasserstein distance of order 2 between λ and ν by*

$$\mathcal{W}_2^2(\lambda,\nu) := \inf_{\zeta\in\Gamma(\lambda,\nu)} \int \|x-y\|^2 \zeta(dx,dy), \tag{S15}$$

where $\Gamma(\lambda,\nu)$ is the set of probability measures ζ on $\mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d)$ satisfying for all $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $\zeta(\mathsf{A}\times\mathbb{R}^d) = \nu(\mathsf{A})$, $\zeta(\mathbb{R}^d \times \mathsf{A}) = \lambda(\mathsf{A})$.

Appendix E.4.2. Proof of the first point in Theorem 3

We prove the following proposition:

Proposition S7. *Under Assumptions 1 to 5, for any linear compression operator $\mathcal{C}$, for any variant v of the algorithm, there exists a limit distribution $\Pi_{\gamma,v}$, which is stationary, such that for any k in $\mathbb{N}$, for any γ satisfying conditions given in Theorems S4 and S5:*

$$\mathcal{W}_2^2((\delta_{w_0} \otimes \otimes_{i=1}^N \delta_{\tau h^i_0}) R^k_\gamma, \Pi_{\gamma,v}) \leq (1-\gamma\mu)^k \int_{(w',h')\in\mathbb{R}^{d(1+N)}} \left\|(w_0, \tau(h^i_0)_i) - (w', \tau(h^i)'_i)\right\|^2 \mathrm{d}\Pi_{\gamma,v}(w', (h^i)'_i).$$

Point 1 in Theorem 3 is derived from the proposition above using $\pi_{\gamma,v} = \pi_{\gamma,v,w}$, with $\pi_{\gamma,v,w}$ as in Equation (S14), the limit distribution of the main iterates $(w_{k-1})_{k\in\mathbb{N}}$ and the observation that:

$$\begin{aligned} \mathcal{W}_2^2(\Theta_k, \pi_{\gamma,v}) &\leq \mathcal{W}_2^2((\delta_{w_0} \otimes \otimes_{i=1}^N \delta_{\tau h^i_0}) R^k_{\gamma,v}, \Pi_{\gamma,v}) \\ &\leq (1-\gamma\mu)^k \int_{(w',h')\in\mathbb{R}^{d(1+N)}} \left\|(w_0, \tau(h^i_0)_i) - (w', \tau(h^i)'_i)\right\|^2 \mathrm{d}\Pi_{\gamma,v}(w', (h^i)'_i) \\ &= (1-\gamma\mu)^k C_0. \end{aligned}$$

The sketch of the proof is simple:

- We introduce a *coupling of random variables* following respectively $\nu_0^a R_{\gamma,v}^k$ and $\nu_0^b R_{\gamma,v}^k$, and show that under the assumptions given in the proposition:

$$\mathcal{W}_2^2(\nu_0^a R_{\gamma,v}^k, \nu_0^b R_{\gamma,v}^k) \leq (1-\gamma\mu)\mathcal{W}_2^2(\nu_0^a R_{\gamma,v}^{k-1}, \nu_0^b R_{\gamma,v}^{k-1}).$$

  This proof follows the same line as the proof of Theorems S4 and S5.

- We deduce that $((\delta_{w_0} \otimes \otimes_{i=1}^N \delta_{\tau h_0^i}) R_{\gamma,v}^k)$ is a Cauchy sequence in a Polish space, thus the existence and stability of the limit, we show that this limit is independent from $(\delta_{w_0} \otimes \otimes_{i=1}^N \delta_{\tau h_0^i}))$ and conclude.

Proof We consider two initial distributions ν_0^a and ν_0^b for $(w_0, \tau(h_0^i)_i)$ with finite second moment and $\gamma > 0$. Let $(w_0^a, \tau(h_0^{i,a})_i)$ and $(w_0^b, \tau(h_0^{i,b})_i)$ be respectively distributed according to ν_0^a and ν_0^b. Let $(w_k^a, \tau(h_k^{i,a})_i)_{k\geq 0}$ and $(w_k^b, \tau(h_k^{i,b})_i)_{k\geq 0}$ the `Artemis` iterates, respectively starting from $(w_0^a, \tau(h_0^{i,a})_i)$ and $(w_0^b, \tau(h_0^{i,b})_i)$, and *sharing the same sequence of noises*, i.e.,

- built with the same gradient oracles $g_k^{i,a} = g_k^{i,b}$ for all $k \in \mathbb{N}, i \in [\![1, N]\!]$.
- the compression operator used for both recursions is almost surely the same, for any iteration k, and both uplink and downlink compression. We denote these operators $\mathcal{C}_{\text{dwn},k}$ and $\mathcal{C}_{\text{up},k}$ the compression operators at iteration k for respectively the uplink compression and downlink compression.

We thus have the following updates, for any $u \in \{a, b\}$:

$$\left\{ \begin{array}{lll} & w_k^u & = w_{k-1}^u - \gamma \mathcal{C}_{\text{dwn},k}\left(\frac{1}{N}\sum_{i=1}^N \mathcal{C}_{\text{up},k}\left(g_k^i - h_{k-1}^{i,u}\right) + h_{k-1}^{i,u}\right) \\ \forall i \in [\![1;n]\!] & h_k^{i,u} & = h_{k-1}^{i,u} + \alpha_{\text{up}} \mathcal{C}_{\text{up},k}\left(g_k^i - h_{k-1}^{i,u}\right). \end{array} \right. \tag{S16}$$

The proof is obtained by induction. For a k in $\mathbb{N}$, let $\left((w_k^a, \tau(h_k^{i,a})_i), (w_k^b, \tau(h_k^{i,b})_i)\right)$ be a coupling of random variable in $\Gamma(\nu_0^a R_{\gamma,v}^k, \nu_0^b R_{\gamma,v}^k)$ – as in Definition 4 –, that achieve the equality in the definition, i.e.,

$$\mathcal{W}_2^2(\nu_0^a R_{\gamma,v}^k, \nu_0^b R_{\gamma,v}^k) = \mathbb{E}\left[\left\|(w_k^a, \tau(h_k^{i,a})_i) - (w_k^b, \tau(h_k^{i,b})_i)\right\|^2\right]. \tag{S17}$$

Existence of such a couple is given by [43, theorem 4.1]. Then $\left((w_k^a, \tau(h_k^{i,a})_i), (w_k^b, \tau(h_k^{i,b})_i)\right)$ obtained after one update from Equation (S16) belongs to $\Gamma(\nu_0^a R_{\gamma,v}^k, \nu_0^b R_{\gamma,v}^k)$, and as a consequence:

$$\begin{aligned} \mathcal{W}_2^2(\nu_0^a R_{\gamma,v}^k, \nu_0^b R_{\gamma,v}^k) &\leq \mathbb{E}\left[\left\|(w_k^a, \tau(h_k^{i,a})_i) - (w_k^b, \tau(h_k^{i,b})_i))\right\|^2\right] \\ &= \mathbb{E}\left[\left\|w_k^a - w_k^b\right\|^2\right] + \tau^2 \sum_{i=1}^N \mathbb{E}\left[\left\|h_k^{i,a} - h_k^{i,b}\right\|^2\right] \\ &= \mathbb{E}\left[\left\|w_k^a - w_k^b\right\|^2\right] + 2\gamma^2 \frac{C}{N} \sum_{i=1}^N \mathbb{E}\left[\left\|h_k^{i,a} - h_k^{i,b}\right\|^2\right], \end{aligned}$$

with $\tau^2 = 2\gamma^2 \frac{C}{N}$, where C depends on the variant as in Theorem 1. We now follow the proof of the previous theorems to control respectively $\mathbb{E}[\|w_k^a - w_k^b\|^2]$ and $\mathbb{E}[\|h_k^{i,a} - h_k^{i,b}\|^2]$. First, following the proof of Equation (S5), we get, using the fact that the compression operator is linear, thus that $\mathcal{C}(x) - \mathcal{C}(y) = \mathcal{C}(x-y)$:

$$\begin{aligned} \mathbb{E}\left[\left\|w_k^a - w_k^b\right\|^2 | \mathcal{H}_{k-1}\right] \leq \left\|w_k^a - w_k^b\right\|^2 &- 2\gamma\left\langle \nabla F(w_{k-1}^a) - \nabla F(w_{k-1}^b), w_k^a - w_k^b\right\rangle \\ &+ \frac{2(2\omega_{\text{up}}+1)(\omega_{\text{dwn}}+1)\gamma^2}{N^2} \sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i(w_{k-1}^a) - g_k^i(w_{k-1}^b)\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \\ &+ \frac{2\omega_{\text{up}}(\omega_{\text{dwn}}+1)\gamma^2}{N^2} \sum_{i=1}^N \mathbb{E}\left[\left\|h_{k-1}^{i,a} - h_{k-1}^{i,b}\right\|^2 \,\middle|\, \mathcal{H}_{k-1}\right] \\ &+ \gamma^2(\omega_{\text{dwn}}+1)L\left\langle \nabla F(w_{k-1}^a) - \nabla F(w_{k-1}^b), w_k^a - w_k^b\right\rangle. \end{aligned}$$

This expression is nearly the same as in Equation (S5), apart from the constant term depending on σ_*^2 that disappears. Note that with a more general compression operator, for example for quantization, it is not possible to derive such a result. Similarly, we control $\mathbb{E}[\|h_k^{i,a} - h_k^{i,b}\|^2]$ using the same line of proof as for Equation (S6), resulting in:

$$\frac{1}{N^2}\sum_{i=0}^{N}\mathbb{E}\left[\left\|h_k^{a,i} - h_k^{b,i}\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \leq \left(1 + p\left(2\alpha_{\rm up}^2\omega_{\rm up} + 2\alpha_{\rm up}^2 - 3\alpha_{\rm up}\right)\right)\frac{1}{N^2}\sum_{i=0}^{N}\mathbb{E}\left[\left\|h_k^{a,i} - h_k^{b,i}\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right]$$
$$+ 2(2\alpha_{\rm up}^2\omega_{\rm up} + 2\alpha_{\rm up}^2 - \alpha_{\rm up})\frac{1}{N^2}\sum_{i=0}^{N}\mathbb{E}\left[\left\|g_k^i(w_{k-1}^a) - g_k^i(w_{k-1}^b)\right\|^2 \;\middle|\; \mathcal{H}_{k-1}\right].$$

Combining both equations, and using Assumptions 1 and 2 and Equation (S17) we get, under conditions on the learning rates $\alpha_{\rm up}, \gamma$ similar to the ones in Theorems S4 and S5, that

$$\mathcal{W}_2^2(\nu_0^a R_{\gamma,v}^k, \nu_0^b R_{\gamma,v}^k) \leq (1-\gamma\mu)\mathcal{W}_2^2(\nu_0^a R_{\gamma,v}^{k-1}, \nu_0^b R_{\gamma,v}^{k-1}).$$

And by induction:

$$\mathcal{W}_2^2(\nu_0^a R_{\gamma,v}^k, \nu_0^b R_{\gamma,v}^k) \leq (1-\gamma\mu)^k\mathcal{W}_2^2(\nu_0^a, \nu_0^b).$$

■

From the contraction above, it is easy to derive the existence of a unique stationnary limit distribution: we use Picard fixed point theorem, as in [8]. This concludes the proof of Proposition S7.

Appendix E.4.3. Proof of the second point of Theorem 3

To prove the second point, we first detail the complementary assumptions mentioned in the text, then show the convergence to the mean squared distance under the limit distribution, and finally give a lower bound on this quantity.

Complementary assumptions.

To prove the lower bound given by the second point, we need to assume that the constants given in the assumptions are tight, in other words, that corresponding lower bounds exist in Assumptions 3 to 5.

Assumption 6 (Lower bound on noise over stochastic gradients computation)**.** *The noise over stochastic gradients at optimal global point for a mini-batch of size b is* lower bounded. *In other words, there exists a constant $\sigma_* \in \mathbb{R}$, such that for all k in $\mathbb{N}$, for all i in $[\![1, N]\!]$, we have a.s:*

$$\mathbb{E}\left[\|g_{k,*}^i - \nabla F_i(w_*)\|^2 \;\middle|\; \mathcal{H}_{k-1}\right] \geq \frac{\sigma_*^2}{b}.$$

Assumption 7 (Lower bound on local gradient at w_*)**.** *There exists a constant $B \in \mathbb{R}$, s.t.:*

$$\frac{1}{N}\sum_{i=1}^{N}\|\nabla F_i(w_*)\|^2 \geq B^2.$$

Assumption 8 (Lower bound on the compression operator's variance)**.** *There exists a constant $\omega \in \mathbb{R}^*$ such that the compression operators $\mathcal{C}_{\rm up}$ and $\mathcal{C}_{\rm dwn}$ verify the following property:*

$$\forall \Delta \in \mathbb{R}^d, \mathbb{E}[\left\|\mathcal{C}_{\rm up/dwn}(\Delta) - \Delta\right\|^2] = \omega_{\rm up/dwn}\|\Delta\|^2 .$$

This last assumption is valid for sparsification, sketching, rand-h, PP.

Moreover, we also assume some extra regularity on the function. This restricts the regularity of the function beyond Assumption 2 and is a purely technical assumption in order to conduct the detailed asymptotic analysis. It is valid in practice for least-squares or logistic regression.

Assumption 9 (Regularity of the functions)**.** *The function F is also times continuously differentiable with second to fifth uniformly bounded derivatives: for all $k \in \{2, \ldots, 5\}$, $\sup_{w\in\mathbb{R}^d}\|F^{(k)}(w)\| < \infty$.*

Convergence of moments.
We first prove that $\mathbb{E}[\|w_{k-1} - w_*\|^2]$ converges to $\mathbb{E}_{w\sim\pi_{\gamma,v}}[\|w - w_*\|^2]$ as k increases to ∞.

We have that the difference satisfies, for random variables w_{k-1} and w following distributions $\delta_{w_0} R^k_{\gamma,v}$ and $\pi_{\gamma,v}$, and coupled such that they achieve the equality in Equation (S15):

$$\begin{aligned}
\Delta_{\mathbb{E},k-1} :&= \mathbb{E}[\|w_{k-1} - w_*\|^2] - \mathbb{E}_{w\sim\pi_{\gamma,v}}[\|w - w_*\|^2] \\
&= \mathbb{E}_{w_{k-1},w\sim\pi_{\gamma,v}}\left[\|w_{k-1} - w_*\|^2 - \|w - w_*\|^2\right] \\
&= \mathbb{E}_{w_{k-1},w\sim\pi_{\gamma,v}}\left[(\|w_{k-1} - w_*\| - \|w - w_*\|)(\|w_{k-1} - w_*\| + \|w - w_*\|)\right] \\
&\overset{\text{C.S}}{\leq} \left(\mathbb{E}_{w_{k-1},w\sim\pi_{\gamma,v}}\left[(\|w_{k-1} - w_*\| - \|w - w_*\|)^2\right]\mathbb{E}_{w_{k-1},w}\left[(\|w_{k-1} - w_*\| + \|w - w_*\|)^2\right]\right)^{1/2} \\
&\overset{\text{T.I.}}{\leq} \left(\mathbb{E}_{w_{k-1},w\sim\pi_{\gamma,v}}\left[(\|w_{k-1} - w\|)^2\right]\mathbb{E}_{w_{k-1},w\sim\pi_{\gamma,v}}\left[(\|w_{k-1} - w_*\| + \|w - w_*\|)^2\right]\right)^{1/2} \\
&\overset{\text{(i)}}{\leq} \left(\mathbb{E}_{w_{k-1},w\sim\pi_{\gamma,v}}\left[(\|w_{k-1} - w\|)^2\right]2L\right)^{1/2} \\
&\overset{\text{(ii)}}{\leq} \left(\mathcal{W}_2^2(\delta_{w_0}R^{k-1}_{\gamma,v}, \pi_{\gamma,v})2L\right)^{1/2} \\
&\overset{\text{(iii)}}{\to} 0.
\end{aligned}$$

Where we have used Cauchy-Schwarz inequality at line C.S., triangular inequality at line T.I., the fact that the moments are bounded by a constant L at line (i), the fact that the distributions are coupled such that they achieve the equality in Equation (S15) at line (ii), and finally Proposition S7 for the conclusion at line (iii).

Overall, this shows that the mean squared distance (i.e., saturation level) converges to the mean squared distance under the limit distribution.

Evaluation of $\mathbb{E}_{w\sim\pi_{\gamma,v}}\|w - w_*\|^2$.
In this section, we denote $\xi(w_{k-1}, h_{k-1})$ the *global* noise, defined by

$$\xi(w_{k-1}, h_{k-1}) = \nabla F(w_{k-1}) - \mathcal{C}_{\text{dwn}}\left(\frac{1}{N}\sum_{i=1}^N \mathcal{C}_{\text{up}}(g^i_k(w_{k-1}) - h^i_{k-1}) + h^i_{k-1}\right),$$

such that $w_k = w_{k-1} - \gamma\nabla F(w_{k-1}) + \gamma\xi(w_{k-1}, h_{k-1})$. In fact, $(\xi)_{k\in\mathbb{N}*}$ is a zero-centered random field characterizing the stochastic oracle on $\nabla F(\cdot)$.

In the following, we denote $a^{\otimes 2} := aa^T$ the second order moment of a. We define Tr the trace operator and Cov the covariance operator such that $\text{Cov}(\xi(w,h)) = \mathbb{E}\left[(\xi(w,h))^{\otimes 2}\right]$, where the expectation is taken on the randomness of both compressions and the gradient oracle. We make a final technical assumption on the regularity of the covariance matrix.

Assumption 10. *We assume that:*

1. $\text{Cov}(\xi(w,h))$ *is continuously differentiable, and there exists constants* C *and* C' *such that for all* $w, h \in \mathbb{R}^{d(1+N)}$, $\max_{o=1,2,3}\text{Cov}^{(o)}(w,h) \leq C + C'||(w,h) - (w_*,h_*)||^2$.
2. $(\xi(w_*,h_*))$ *has finite order moments up to order 8.*

Remark: with the *linear* operators, this assumption can directly be translated into an assumption on the moments and regularity of g^i_k. Note that Point 2 in Assumption 10 is an extension of Assumption 3 to higher order moments, but **still at the optimal point**. Under this assumption, we have the following lemma:

Lemma S14. *Under Assumptions 1 to 5, 9 and 10, we have that*

$$\mathbb{E}_{\pi_{\gamma,v}}\left[\|w - w_*\|^2\right] \underset{\gamma\to 0}{=} \gamma\text{Tr}\left(A\ \text{Cov}(\xi(w_*,h_*))\right) + O(\gamma^2), \tag{S18}$$

with $A := (F''(w_*)\otimes I + I\otimes F''(w_*))^{-1}$.

The intuition of the proof is natural: using the stability of the limit distribution, we have that if we start from the stationary distribution, i.e., $(w_0,h_0)\sim\Pi_{\gamma,v}$, then $(w_1,h_1)\sim\Pi_{\gamma,v}$.

We can thus write:

$$\begin{aligned}
\mathbb{E}_{\pi_{\gamma,v}}\left[(w - w_*)^{\otimes 2}\right] &= \mathbb{E}\left[(w_1 - w_*)^{\otimes 2}\right] \\
&= \mathbb{E}\left[(w_0 - w_* - \gamma\nabla F(w_0) + \gamma\xi(w_0,h_0))^{\otimes 2}\right].
\end{aligned}$$

Then, expanding the right hand side and using the fact that $\mathbb{E}[\xi(w_0,h_0)|\mathcal{H}_0]=0$, then the fact that $\mathbb{E}\left[(w_1-w_*)^{\otimes 2}\right]=\mathbb{E}\left[(w_0-w_*)^{\otimes 2}\right]$, and expanding the derivative of F around w_* (this is where we require the regularity assumption Assumption 9), we get that:

$$\gamma\left(F''(w_*)\otimes I+I\otimes F''(w_*)+O(\gamma)\right)\mathbb{E}_{\pi_{\gamma,v}}\left[(w-w_*)^{\otimes 2}\right]\underset{\gamma\to 0}{=}\gamma^2\mathbb{E}_{(w,h)\sim\Pi_{\gamma,v}}\left[\xi(w,h)^{\otimes 2}\right].$$

Thus:

$$\mathbb{E}_{\pi_{\gamma,v}}\left[(w-w_*)^{\otimes 2}\right]\underset{\gamma\to 0}{=}\gamma A\mathbb{E}_{(w,h)\sim\Pi_{\gamma,v}}\left[\xi(w,h)^{\otimes 2}\right]+O(\gamma^2).$$
$$\Rightarrow\mathbb{E}_{\pi_{\gamma,v}}\left[\|(w-w_*)\|^2\right]\underset{\gamma\to 0}{=}\gamma\mathrm{Tr}\left(A\mathbb{E}_{(w,h)\sim\Pi_{\gamma,v}}\left[\xi(w,h)^{\otimes 2}\right]\right)+O(\gamma^2).$$

Finally, we use that $\mathbb{E}_{(w,h)\sim\Pi_{\gamma,v}}\left[\mathrm{Cov}(\xi(w,h))\right]\underset{\gamma\to 0}{=}\mathrm{Cov}(\xi(w_*,h_*))+O(\gamma)$ (which is derived from Assumption 10) to get Lemma S14. More formally, we can rely on Theorem 4 in Dieuleveut et al. [8]: under Assumptions 1 to 5 and Assumptions 9 and 10, all assumptions required for the application of the theorem are verified and the result follows.

To conclude the proof, it only remains to control $\mathrm{Cov}(\xi(w_*,h_*))$. We have the following Lemma:

Lemma S15. *Under Assumptions 6 to 8, we have that, for any variant v of the algorithm, with the constant E given in Theorem 1 depending on the variant:*

$$\mathrm{Tr}\left(\mathrm{Cov}(\xi(w_*,h_*))\right)=\Omega\left(\frac{\gamma E}{\mu N}\right). \tag{S19}$$

Combining Lemmas S14 and S15 and using the observation that A is lower bounded by $\frac{1}{2L}$ independently of γ,N,σ_*,B, we have proved the following proposition:

Proposition S8. *Under Assumptions 1 to 5 and 6 to 10, we have that*

$$\mathbb{E}[\|w_{k-1}-w_*\|^2]\underset{k\to\infty}{\to}\mathbb{E}_{\pi_{\gamma,v}}\left[\|w-w_*\|^2\right]\underset{\gamma\to 0}{=}\Omega\left(\frac{\gamma E}{\mu N}\right)+O(\gamma^2), \tag{S20}$$

where the constant in the Ω is independent of N,σ_*,γ,B (it depends only on the regularity of the operator A).

Before giving the proof, we make a couple of observations:

1. This shows that the upper bound on the limit mean squared error given in Theorem 1 **is tight** with respect to N,σ_*,γ,B. This underlines that the conditions on the problem that we have used are the correct ones to understand convergence.
2. The upper bound is possibly not tight with respect to μ, as is clear from the proof: the tight bound is actually $\mathrm{Tr}\left(A\mathrm{Cov}(\xi(w_*,h_*))\right)$. Getting a tight upper bound involving the eigenvalue decomposition of A instead of only μ is an open direction.
3. In the memory-less case, $h\equiv 0$ and all the proof can be carried out analyzing only the distribution of the iterates $(w_{k-1})_k$ and not necessarily the couple $(w_{k-1},(h^i_{k-1})_i)_k$.

We now give the proof of Lemma S15.

Proof With memory, we have the following:

$$\begin{aligned}
\mathrm{Tr}\left(\mathrm{Cov}(\xi(w_*,h_*))\right)&=\mathbb{E}\left[\left\|\mathcal{C}_{\mathrm{dwn}}\left(\frac{1}{N}\sum_{i=1}^N\mathcal{C}_{\mathrm{up}}(g_1^i(w_*)-h_*^i)+h_*^i\right)\right\|^2\right]\\
&\overset{\text{(i)}}{=}(1+\omega_{\mathrm{dwn}})\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N\mathcal{C}_{\mathrm{up}}(g_1^i(w_*)-h_*^i)+h_*^i\right\|^2\right]\\
&\overset{\text{(ii)}}{=}\frac{(1+\omega_{\mathrm{dwn}})}{N^2}\sum_{i=1}^N\mathbb{E}\left[\left\|\mathcal{C}_{\mathrm{up}}(g_1^i(w_*)-h_*^i)\right\|^2\right]\\
&\overset{\text{(iii)}}{=}\frac{(1+\omega_{\mathrm{dwn}})}{N^2}\sum_{i=1}^N(1+\omega_{\mathrm{up}})\mathbb{E}\left[\left\|g_1^i(w_*)-h_*^i\right\|^2\right]\\
&\overset{\text{(iv)}}{\geq}\frac{(1+\omega_{\mathrm{dwn}})}{N}(1+\omega_{\mathrm{up}})\frac{\sigma_*^2}{b}.
\end{aligned}$$

At line (i) we use Assumption 8 for the downlink compression operator with constant ω_{dwn}. At line (ii) we use the fact that $\sum_{i=1}^N h_*^i = \nabla F(w_*) = 0$, the independence of the random variables $\mathcal{C}_{\text{up}}(g_1^i(w_*) - h_*^i), \mathcal{C}_{\text{up}}(g_1^j(w_*) - h_*^j)$ for $i \neq j$ and the fact that they have 0 mean. We use Assumption 8 for the uplink compression operator with constant ω_{up} in line (iii); and finally Assumption 6 at line (iv) to lower bound the variance of the gradients at the optimum. This proof applies to both simple and double compression with $\omega_{\text{dwn}} = 0$ or not.

Remark that for the variant 2 of `Artemis`, the constant E given in Theorem 1 has a factor $\alpha_{\text{up}}^2 C(\omega+1)$: combining with the value of C, this term is indeed of the order of $(1+\omega_{\text{dwn}})(1+\omega_{\text{up}})$.

Without memory, we have the following computation:

$$
\begin{aligned}
\text{Tr}\left(\text{Cov}(\xi(w_*, 0))\right) &= \mathbb{E}\left[\left\|\mathcal{C}_{\text{dwn}}\left(\frac{1}{N}\sum_{i=1}^N \mathcal{C}_{\text{up}}(g_1^i(w_*))\right)\right\|^2\right] \\
&\overset{\text{(i)}}{=} (1+\omega_{\text{dwn}})\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \mathcal{C}_{\text{up}}(g_1^i(w_*)) - h_*^i\right\|^2\right] \\
&\overset{\text{(ii)}}{=} \frac{(1+\omega_{\text{dwn}})}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|\mathcal{C}_{\text{up}}(g_1^i(w_*)) - h_*^i\right\|^2\right] \\
&\overset{\text{(iii)}}{=} \frac{(1+\omega_{\text{dwn}})}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|\mathcal{C}_{\text{up}}(g_1^i(w_*)) - g_1^i(w_*)\right\|^2 + \left\|g_1^i(w_*) - h_*^i\right\|^2\right]
\end{aligned}
$$

At line (i) we use Assumption 8 for the downlink compression operator with constant ω_{dwn} and the fact that $\sum_{i=1}^N h_*^i = \nabla F(w_*) = 0$, then at line (ii) the independence of the random variables $\mathcal{C}_{\text{up}}(g_1^i(w_*)) - h_*^i$ with mean 0, then a Bias Variance decomposition at line (iii).

$$
\begin{aligned}
\text{Tr}\left(\text{Cov}(\xi(w_*, 0))\right) &\overset{\text{(iv)}}{=} \frac{(1+\omega_{\text{dwn}})}{N^2}\sum_{i=1}^N \mathbb{E}\left[\omega_{\text{up}}\left\|(g_1^i(w_*))\right\|^2 + \left\|g_1^i(w_*) - h_*^i\right\|^2\right] \\
&\overset{\text{(v)}}{=} \frac{(1+\omega_{\text{dwn}})}{N^2}\sum_{i=1}^N \mathbb{E}\left[\omega_{\text{up}}\left(\left\|g_1^i(w_*) - h_i^*\right\|^2 + \|h_i^*\|^2\right) + \left\|g_1^i(w_*) - h_*^i\right\|^2\right] \\
&\overset{\text{(vi)}}{=} \frac{(1+\omega_{\text{dwn}})}{N}\left((\omega_{\text{up}}+1)\frac{\sigma_*^2}{b} + \omega_{\text{up}}B^2\right).
\end{aligned}
$$

Next we use Assumption 8 for the uplink compression operator with constant ω_{up} at line (iv). Line (v) is another Bias-Variance decomposition and we finally conclude by using Assumptions 6 and 7 at line (vi) and reorganizing terms.

We have showed the lower bound both with or without memory, which concludes the proof. ■