Artemis: tight convergence guarantees for bidirectional compression in Federated Learning

Constantin Philippenko **Aymeric Dieuleveut**
CMAP, École Polytechnique, Institut Polytechnique de Paris
[fistname].[lastname]@polytechnique.edu

Abstract

We introduce a new algorithm - Artemis - tackling the problem of learning in a distributed framework with communication constraints. Several workers (randomly sampled) perform the optimization process using a central server to aggregate their computation. To alleviate the communication cost, Artemis compresses the information sent in *both directions* (from the workers to the server and conversely) combined with a memory mechanism. It improves on existing quantized federated learning algorithms that only consider unidirectional compression (to the server), or use very strong assumptions on the compression operator, and often do not take into account devices partial participation. 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, analyze Polyak-Ruppert averaging. We use convergence in distribution to obtain a *lower bound* of the asymptotic variance that highlights practical limits of compression. And we provide experimental results to demonstrate the validity of our analysis.

1 Introduction

In modern large scale machine learning applications, optimization has to be processed in a distributed fashion, using a potentially large number N of workers. In the data-parallel framework, each worker only accesses a fraction of the data: new challenges have arisen, especially when communication constraints between the workers are present, or when the distributions of the observations on each node are different. These challenges are tackled in the Federated Learning (FL) setting, introduced by Konečný et al. [2016] and McMahan et al. [2017]. Kairouz et al. [2019] recently gave a complete overview of the field and described open problems, underlining the importance of the communication constraint.

In this paper, we focus on first-order methods, especially Stochastic Gradient Descent [Bottou, 1999; Robbins and Monro, 1951] in a centralized framework: a central machine aggregates the computation of the N workers in a synchronized way. 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. Several papers have considered this problem. However, most of them [Karimireddy et al., 2019; Mishchenko et al., 2019; Alistarh et al., 2017; Agarwal et al., 2018; Li et al., 2020; Wu et al., 2018; Horváth and Richtárik, 2020; Horváth et al., 2019] focus on compressing the message sent from the workers to the central node. As the central node also needs to broadcast the message back to the workers, the global communication budget can only be cut by a mere factor of 2. In this paper, we focus on compressing the message sent *in both directions*.

Of course, 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 Report, 2020]. Results are provided in Figure 1, comparing download and upload speeds (see Appendix B for details). 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 $s = 1$ (see Appendix A.2), 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.

Double compression has been recently considered by Tang et al. [2019], but theoretical guarantees were only

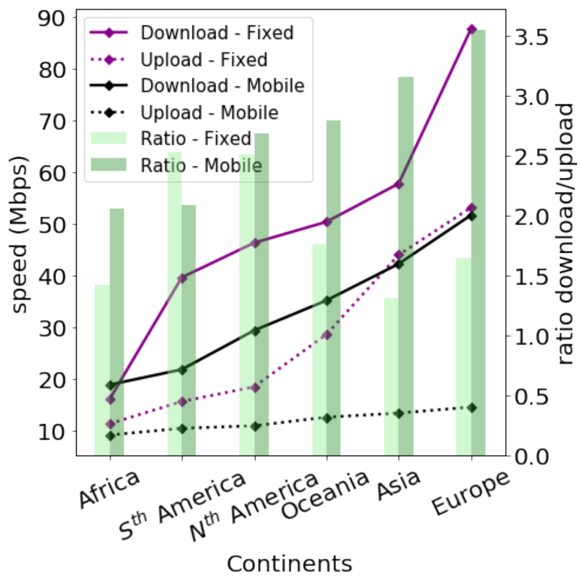


Figure 1: Left axis: upload & download speed for mobile and fixed broadband. Right axis: ratio (green bars). The dataset is gathered from Report [2020].

provided under a bounded error on the compression, which is neither realistic nor tight for quantization, which is one of the most widely used unbiased compression operator. Several extensions have been proposed in [Zheng et al., 2019; Liu et al., 2020; Yu et al., 2019]. The most recent, Dore, defined by Liu et al. [2020], analyzed a double compression approach, combining error compensation, memory mechanism and model compression, with a uniform bound on the gradient variance. In this work, we provide new results on Dore-like algorithms, considering a framework *without error-feedback*, combined with a partial participation of devices, using tighter assumptions and quantifying precisely the impact of data heterogeneity on the convergence.

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:

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

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 $(z_k^i)_{1 \leq k \leq n_i}$ 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 worker 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).

Assumptions made on the oracle directly influence the convergence rate of the algorithm: in this paper, we neither assume that the gradients are uniformly bounded [as in Zheng et al., 2019] nor that their variance is uniformly bounded [as in Alistarh et al., 2017; Mishchenko et al., 2019; Liu et al., 2020; Tang et al., 2019; Horváth et al., 2019]: 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 [Dieuleveut et al., 2019; Gower et al., 2019] 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 [Belkin et al., 2019]. 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., Zhang et al., 2017] or in large dimension regression [Mei and Montanari, 2019].

Moreover, we consider the non independent and identically distributed (non i.i.d.) setting: we assume that the distribution of the data on each worker can be different, and *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 use a "memory" process [Mishchenko et al., 2019].

Overall, we make the following contributions:

1. We describe a new algorithm – Artemis – which uses a bidirectional compression with memory and take into account partial participation of devices. We provide and analyze in Theorem 1 a fast rate of convergence – exponential convergence up to a threshold proportional to σ_*^2 – for Artemis and two of its variants that match earlier algorithms QSGD and Diana, obtaining tighter bounds for these particular algorithms than in respectively [Alistarh et al., 2017] and [Mishchenko et al., 2019].
2. We explicitly tackle non i.i.d. data using Assumption 4, proving that the limit variance of Artemis 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 convergence of Artemis using a Polyak-Ruppert averaging in Theorem 2, underlining the impact of memory in the non i.i.d. setting.

4. We prove *convergence in distribution* of the iterates in Theorem 3, 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.

Furthermore, we support our analysis with various experiments illustrating the behavior of our new algorithm and we provide the code to reproduce our experiments on `Artemis`. See this repository. 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 algorithm mechanism. 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 and 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 denote w_* the optimal parameter and $h_*^i = \nabla F_i(w_*)$, for i in $[\![1, N]\!]$. We use $\|\cdot\|$ to denote the Euclidean norm. To solve this problem, we rely on a stochastic gradient descent (SGD) algorithm.

A stochastic gradient g_{k+1}^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: to alleviate notation, we will use $g_{k+1}^i = g_{k+1}^i(w_k)$ and $g_{k+1,*}^i = g_{k+1}^i(w_*)$ to denote the stochastic gradient vectors at points w_k and w_* on device i. In the classical centralized framework (without compression), with partial participation of devices, SGD corresponds to:

$$w_{k+1} = w_k - \gamma \frac{1}{pN} \sum_{i \in S_k} g_{k+1}^i = w_k - \gamma g_{k+1,S_k} \,, \quad (2)$$

where γ is the learning rate, S_k is the set of *active* (i.e., participating) workers at step k and p is the probability that each device participates.

However, computing such a sequence would require the nodes to send either the gradient g_{k+1}^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+1,S_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 [Mishchenko et al., 2019; Li et al., 2020]. 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. [2019], we always broadcast gradients and never models. To distinguish the two compression operations we denote $\mathcal{C}_{\mathrm{up}}$ and $\mathcal{C}_{\mathrm{dwn}}$ the compression operator for downlink and uplink. 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: $\widehat{\Delta}_k^i = \mathcal{C}_{\mathrm{up}}(g_{k+1}^i - h_k^i)$, and updates the *memory term* $h_{k+1}^i = h_k^i + \alpha \widehat{\Delta}_k^i$ with $\alpha \in \mathbb{R}^*$. 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: $\Omega_{k+1,S_k} = \mathcal{C}_{\mathrm{dwn}}\left(\frac{1}{pN}\sum_{i\in S_k} \widehat{\Delta}_k^i + h_k^i\right)$. 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:

$$\begin{cases} \forall i \in [\![1, N]\!] \,, \quad \widehat{\Delta}_k^i = \mathcal{C}_{\mathrm{up}}\left(g_{k+1}^i - h_k^i\right) \\ \Omega_{k+1} = \mathcal{C}_{\mathrm{dwn}}\left(\frac{1}{pN}\sum_{i\in S_k}(\widehat{\Delta}_k^i + h_k^i)\right). \\ w_{k+1} = w_k - \gamma \Omega_{k+1,S_k} \end{cases} \quad (3)$$

Constants $\gamma, \alpha \in \mathbb{R}^*$ are learning rates for respectively the iterate sequence and the memory sequence. This is illustrated on Algorithm 1 and fig. 2.

Because un-active devices are not updated at steps to which they do not participate, they have to catch up the sequence of $(\Omega_k)_{k\in\mathbb{N}}$, in order to evaluate the gradient on the same up-to-date model than other workers. This mechanism is described in the pseudo-code in Algorithm 1 and does not interfere with the theoretical analysis. Remark that we have here used the *true value* of p in the update (to get an unbiased estimator of the gradient), but it is obviously possible to use an estimated value $\hat{p}$ in Equation (3): indeed, it is exactly equivalent to multiplying the step size γ by a factor $p/\hat{p}$, thus neither changes the practical implementation nor the theoretical analysis.

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

2.1 Assumptions

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

Assumption 1 (Strong convexity)**.** *F is μ-strongly*

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 Alistarh et al. [2017] for QSGD, Mishchenko et al. [2019] for Diana, Horváth and Richtárik [2020] for [HR20], Liu et al. [2020] for Dore and Tang et al. [2019] for DoubleSquezze.

	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 compensation			✓	✓	✓	✓	
Memory		✓		✓			✓
Device sampling			✓				✓

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 $\mathcal{C}_{\mathrm{up}}$ and $\mathcal{C}_{\mathrm{dwn}}$.
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.**
  - Send $(\Omega_j)_{j=k_i+1}^{k}$ 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 $g_{k+1}^i = g_{k+1}(w_k)$ (with mini-batch)
  - Set $\Delta_k^i = g_{k+1}^i - h_k^i$, compress it $\widehat{\Delta}_k^i = \mathcal{C}_{\mathrm{up}}(\Delta_k^i)$
  - Update memory term: $h_{k+1}^i = h_k^i + \alpha \widehat{\Delta}_k^i$
  - Send $\widehat{\Delta}_k^i$ to central server
- Compute $\widehat{g}_{k+1, S_k} = \dfrac{1}{pN} \sum_{i \in S_k} \widehat{\Delta}_k^i + h_k^i$
- Update memory terms kept on central server: $\forall i \in S_k, h_{k+1}^i = h_k^i + \alpha \widehat{\Delta}_k^i$
- Back compression: $\Omega_{k+1, S_k} = G_{k+1, S_k} = \mathcal{C}_{\mathrm{dwn}}(\widehat{g}_{k+1, S_k})$
- Broadcast Ω_{k+1} to all workers.
- Update model on central server: $w_{k+1} = w_k - \gamma \Omega_{k+1, S_k}$

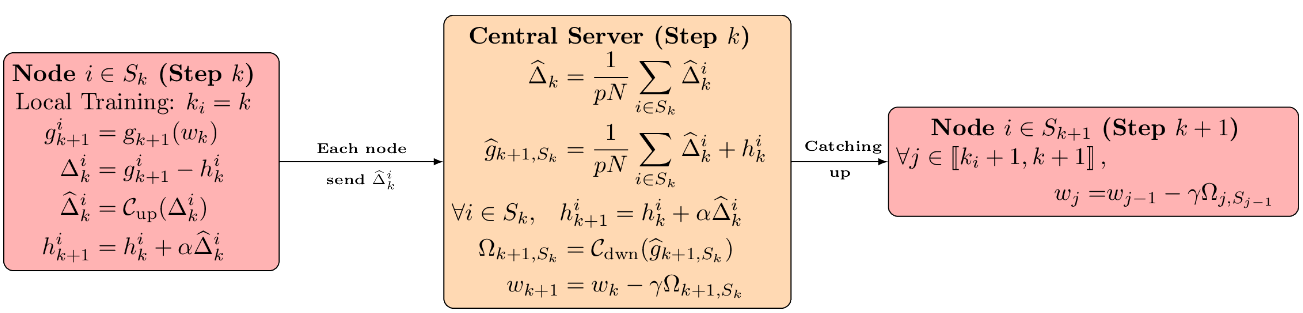


Figure 2: Visual illustration of Artemis bidirectional compression at iteration $k \in \mathbb{N}$. w_k is the model's parameter, $\mathcal{C}_{\mathrm{up}}$ and $\mathcal{C}_{\mathrm{dwn}}$ are the compression operators, γ is the step size, α is the learning rate which simulates a memory mechanism of past iterates and allows the compressed values to tend to zero.

convex, that is for all vectors w, v in $\mathbb{R}^d$:

$$F(v) \geq F(w) + (v-w)^T \nabla F(w) + \frac{\mu}{2}\|v-w\|_2^2 .$$

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 $v = w_*$ in the proof of Theorems 1 and 2.

Below, we introduce cocoercivity [see Zhu and Marcotte, 1996, for more details about this hypothesis]. This assumption implies that all $(F_i)_{i \in [\![1,N]\!]}$ are L-smooth. In machine learning framework, this is for example valid for both Least Squares Regression (LSR) and Logistic Regression (LR) [Patel and Dieuleveut, 2019].

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 w_1, w_2 in $\mathbb{R}^d$, we have:*

$$\mathbb{E}[\|g_k^i(w_1) - g_k^i(w_2)\|^2] \leq L \left\langle \nabla F_i(w_1) - \nabla F_i(w_2) \mid w_1 - w_2 \right\rangle .$$

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.:

$$\|g_k^i(w_1) - g_k^i(w_2)\|^2 \leq L \left\langle g_k^i(w_1) - g_k^i(w_2) \mid w_1 - w_2 \right\rangle .$$

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.3.

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 $\sigma_* \in \mathbb{R}$, such that for all k in $\mathbb{N}$, for all i in $[\![1,N]\!]$, we have a.s:*

$$\mathbb{E}\left[\|g_{k+1,*}^i - \nabla F_i(w_*)\|^2 \,\Big|\, \mathcal{H}_k\right] \leq \frac{\sigma_*^2}{b}.$$

The constant σ_*^2 is null, e.g. 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` [Mishchenko et al., 2019; Li et al., 2020], `Dore` [Liu et al., 2020], `Dist-EF-SGD` [Zheng et al., 2019] or `Double-Squeeze` [Tang et al., 2019], 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 variance is null ($\sigma_*^2 = 0$) at 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 LSR setting, while the uniform bound *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.:*

$$\forall i \in [\![1,N]\!], \quad \frac{1}{N}\sum_{i=0}^{N} \|\nabla F_i(w_*)\|^2 \leq B^2 .$$

This assumption is used to quantify how different the distributions are on the different machines. 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 $\mathbb{N}$, a.s., $\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\|g_{k+1,*}^i\|^2 \,\Big|\, \mathcal{H}_k\right] \leq \frac{\sigma_*^2}{b} + B^2$.

Finally, compression operators can be classified in two main categories: quantization [as in Alistarh et al., 2017; Seide et al., 2014; Zhou et al., 2018; Wen et al., 2017; Reisizadeh et al., 2019; Horváth et al., 2019] and sparsification [as in Aji and Heafield, 2017; Alistarh et al., 2018; Khirirat et al., 2020]. 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 $\mathcal{C}_{\mathrm{up}}$ and $\mathcal{C}_{\mathrm{dwn}}$:

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

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

In other words, the compression operators are unbiased and their variances are bounded. Note that Horváth and Richtárik [2020] have shown that using an unbiased operator leads to better performances. Unlike us, Tang et al. [2019] assume uniformly bounded compression error, which is a much more restrictive assumption.

We also encompass partial device participation.

Assumption 6. *At each round k in $\mathbb{N}$, each device has a probability p of participating, independently from other workers, i.e., there exists a sequence $(B_k^i)_{k,i}$ of i.i.d. Bernoulli random variables $\mathcal{B}(p)$, such that for any k and i, B_k^i marks if device i is active at step k. We denote $S_k = \{i \in [\![1,N]\!] / B_k^i = 1\}$ the set of active devices at round k and $N_k = \mathrm{card}(S_k)$ the number of active workers.*

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. [2014] 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. [2019]. Alistarh et al. [2017] 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. [2018] add an error compensation mechanism which accumulates quantization errors and corrects the gradient computation at each iteration. `Diana` [introduced in Mishchenko et al., 2019] introduces a "memory" term in the place of accumulating errors. Li et al. [2020] extend this algorithm and improve its convergence by using an accelerated gradient descent. Reisizadeh et al. [2019] 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 and Richtárik [2020] detail results that also consider devices partial participation. Tang et al. [2019] are the 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. Bidirectional compression is later developed only in Yu et al. [2019]; Liu et al. [2020]; Zheng et al. [2019]. Instead of compressing gradients, Yu et al. [2019] choose to compress models. This approach is enhanced by Liu et al. [2020] who combine model compression with a memory mechanism and an error compensation drawing from Mishchenko et al. [2019]. Both Tang et al. [2019] and Zheng et al. [2019] compress gradients without using a memory mechanism. However, as proved in the following section, memory is key to reducing the asymptotic variance in the non i.i.d. case.

Variants of `Artemis` and link with other algorithms: Note that $\omega_{\mathcal{C}}^{\mathrm{up/dwn}}$ are considered as *parameters* of the algorithm, as the compression level can be chosen, as well as γ and α (while other constants B, σ_*, p, μ, L are truly "assumptions" on the data distribution, oracle, and activation probabilities). Having $\omega_{\mathcal{C}}^{\mathrm{up/dwn}} = 0$ is equivalent to not compressing the signal. We can thus define different *variants* of `Artemis`: the variants with unidirectional compression ($\omega_{\mathcal{C}}^{\mathrm{dwn}} = 0$) w.o. or with memory ($\alpha = 0$ or $\alpha \neq 0$) recover `QSGD` and `DIANA`. The variant using bidirectional compression ($\omega_{\mathcal{C}}^{\mathrm{dwn}} \neq 0$) w.o memory ($\alpha \neq 0$) is called `Bi-QSGD`. The last variant, using bidirectional compression *and* memory is the one we refer to as `Artemis` if no precision is given.

We now provide theoretical results about the convergence of bidirectional compression.

3 Theoretical results

In this section, we present our main theoretical results on the convergence of `Artemis` and its variants. For the sake of clarity, the most complete and tightest versions of theorems are given in Appendices, and simplified versions are provided here. The main linear convergence rates are given in Theorem 1. In Theorem 2 we show that `Artemis` combined with Polyak-Ruppert averaging reaches a sub-linear convergence rate. In this section, we denote $\delta_0^2 = \|w_0 - w_*\|^2$.

Theorem 1 (Convergence of `Artemis`). *Under Assumptions 1 to 6, for a step size γ satisfying the conditions in Table 3, for a learning rate α 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:*

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

for constants C and E depending on the variant (independent of k) given in Table 2 or in the appendix. Variants with $\alpha \neq 0$ require $\alpha \in [1/2(\omega_{\mathcal{C}}^{\mathrm{up}} + 1), \alpha_{\max}]$, 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. $\alpha \neq 0$) it also depends on the difference between distributions (constant B^2).
3. **Variance term and memory.** On the other hand, the variance depends a) on both σ_*^2/b, and the distributions' difference B^2 without memory b) only on the gradients' variance *at the optimum* σ_*^2/b with memory. Similar theorems in related literature Liu et al. [2020]; Alistarh et al. [2017]; Mishchenko et al. [2019]; Yu et al. [2019]; Tang et al. [2019]; Zheng et al. [2019] 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 paper

Table 2: Details on constants C and E defined in Theorem 1.

α	E	C
0	$(\omega_{\mathcal{C}}^{\text{dwn}}+1)\left((\omega_{\mathcal{C}}^{\text{up}}+1)\frac{\sigma_*^2}{b}+(\omega_{\mathcal{C}}^{\text{up}}+1-p)B^2\right)$	$C=0$
$\neq 0$	$\frac{\sigma_*^2}{b}\left((2\omega_{\mathcal{C}}^{\text{up}}+1)(\omega_{\mathcal{C}}^{\text{dwn}}+1)+4\alpha^2C(\omega_{\mathcal{C}}^{\text{up}}+1)-2\alpha C\right)$	$C>0$ (see Th. S5)

and [Liu et al., 2020] 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 $\omega_{\mathcal{C}}^{\text{up}}$ for the unidirectional compression, and proportional to $\omega_{\mathcal{C}}^{\text{up}} \times \omega_{\mathcal{C}}^{\text{dwn}}$ 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 [Mishchenko et al., 2019] and [Liu et al., 2020]. 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 $\alpha=0$, if $\omega_{\mathcal{C}}^{\text{up/dwn}}=0$ (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: $pN \gg \omega_{\mathcal{C}}^{\text{up}}$, $pN \approx \omega_{\mathcal{C}}^{\text{up}}$ and $\omega_{\mathcal{C}}^{\text{up}} \gg pN$. Using bidirectional compression impacts $\gamma_{\max}$ by a factor $\omega_{\mathcal{C}}^{\text{dwn}}+1$ in comparison to unidirectional compression. For unidirectional compression, if the number of machines is at least of the order of $\omega_{\mathcal{C}}^{\text{up}}$, then $\gamma_{\max}$ nearly corresponds to $\gamma_{\max}$ for vanilla (serial) SGD.

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 6 (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 $\gamma=\min\left(\sqrt{\frac{pN\delta_0^2}{2EK}};\gamma_{\max}\right)$,*

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

Memory	$\alpha=0$	$\alpha\neq 0$
$pN \gg \omega_{\mathcal{C}}^{\text{up}}$	$\frac{1}{(\omega_{\mathcal{C}}^{\text{dwn}}+1)L}$	$\frac{1}{2(\omega_{\mathcal{C}}^{\text{dwn}}+1)L}$
$pN \approx \omega_{\mathcal{C}}^{\text{up}}$	$\frac{1}{3(\omega_{\mathcal{C}}^{\text{dwn}}+1)L}$	$\frac{1}{5(\omega_{\mathcal{C}}^{\text{dwn}}+1)L}$
$\omega_{\mathcal{C}}^{\text{up}} \gg pN$	$\frac{pN}{2\omega_{\mathcal{C}}^{\text{up}}(\omega_{\mathcal{C}}^{\text{dwn}}+1)L}$	$\frac{pN}{4\omega_{\mathcal{C}}^{\text{up}}(\omega_{\mathcal{C}}^{\text{dwn}}+1)L}$

with $\gamma_{\max}$ as in Table 3, we have a sublinear convergence rate for the Polyak-Ruppert averaged iterate $\bar{w}_k=\frac{1}{K}\sum_{k=0}^{K} w_k$, with $\varepsilon_F(\bar{w}_k)=F(\bar{w}_k)-F(w_*)$:

$$\varepsilon_F(w_k) \leq 2\max\left(\sqrt{\frac{2\delta_0^2E}{pNK}};\frac{\delta_0^2}{\gamma_{\max}K}\right)+\frac{2\gamma_{\max}CB^2}{p^2K}.$$

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 [Rakhlin et al., 2012].

2. **Speed of convergence, if $\sigma_*=0$, $B\neq 0$, $K\to\infty$.** For $\alpha\neq 0$, $E=0$, while for $\alpha=0$, $E\propto B^2$. 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 $E \leq \delta_0^2 pN/(2K\gamma_{\max}^2)$ – 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 $\min\{\omega_{\mathcal{C}}^{\text{up}},\omega_{\mathcal{C}}^{\text{dwn}}\}$-times more iterations, this is no longer beneficial if the convergence rate is dominated by $\sqrt{2\delta_0^2E/pNK}$, as E increases by a factor $\omega_{\mathcal{C}}^{\text{up}}\times\omega_{\mathcal{C}}^{\text{dwn}}$.

4. **Memoryless case, impact of minibatch.** In the variant of `Artemis` *without memory*, the asymptotic convergence rate is $\sqrt{2\delta_0^2E/pNK}$ 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 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 $\mathcal{W}_2$ [Villani, 2009]: this gives us a lower bound on the asymptotic variance. Here, we further assume that the compression operator is *Stochastic sparsification* [Wen et al., 2017].

Theorem 3 (Convergence in distribution and lower bound on the variance)**.** *Under Assumptions 1 to 6, for γ, α, 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 $k \geq 1$, $\mathcal{W}_2(\Theta_k, \pi_{\gamma,v}) \leq (1-\gamma\mu)^k C_0$, with C_0 a constant.*
2. *When k goes to ∞, 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 $\Omega(\gamma E/\mu p N)$ 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. $\sigma_*^2, \omega_{\mathcal{C}}^{\mathrm{up}}, \omega_{\mathcal{C}}^{\mathrm{dwn}}, B^2, N$ and γ. Especially, it proves that there is indeed a quadratic increase in the variance w.r.t. $\omega_{\mathcal{C}}^{\mathrm{up}}$ and $\omega_{\mathcal{C}}^{\mathrm{dwn}}$ 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, this is *the first result* giving a lower bound on the asymptotic variance for algorithms using compression.

Proof and assumptions. This theorem also naturally requires, for the second point, Assumptions 3 to 5 to be "tight": that is, e.g. $\mathrm{Var}(g_{k+1,*}^i) \geq \Omega(\sigma_*^2/b)$; 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. We compare `Artemis` with existing works on compression: `QSGD`, `Diana`, `Bi-QSGD` (which is `QSGD` but with a bidirectional compression) and also usual `SGD` without compression. **We do not compare to Liu et al. [2020] as we focus on algorithms not using the error compensation process**, but `Artemis` can be seen as a modification of `Dore`, and they have the same convergence behavior.

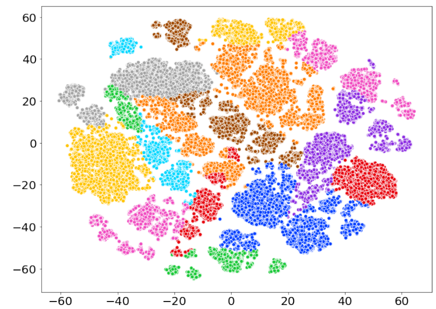


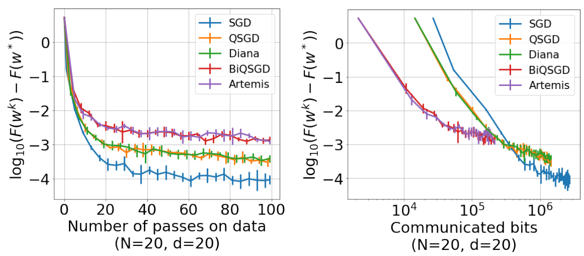
Figure 3: TSNE representation and clusters for *quantum*.

We display the excess error $F(w_t) - F(w_*)$ w.r.t. the number of iterations t or the number of communicated bits. We use a 1-quantization scheme (defined in Appendix A.2 for all experiments; with $s = 1$, which is the most drastic compression).

We first consider two simple synthetic datasets: one for least-square regression (with the same distribution over each machine), and one for logistic regression (with varying distributions across machines); more details are given in Appendix C on the way data is generated. We use $N = 20$ devices, each holding 200 points of dimension $d = 20$, and run algorithms over 100 epochs.

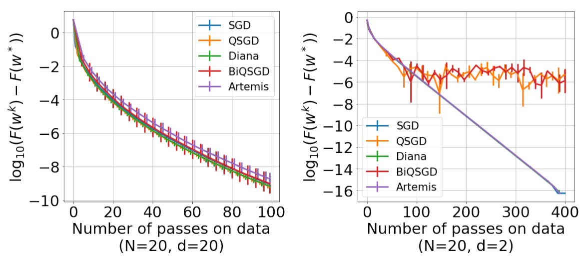
We then considered two real-world dataset to illustrate how Artemis is performing on real data and higher dimension: *superconduct* [see Hamidieh, with 21 263 points and 81 features] and *quantum* [see Caruana et al., with 50 000 points and 65 features] with $N = 10$ workers. To simulate non-i.i.d *and* unbalanced workers, we split the dataset in heterogeneous groups, using a Gaussian mixture clustering on TSNE representations (defined by Maaten and Hinton). Thus, the distribution and number of points hold by each worker largely differs between devices, see Figure 3 for illustration. Curves are averaged over 5 runs, and we plot error bars on all figures. To compute error bars we use the standard deviation of the logarithmic difference between the loss function at iteration k and the objective loss, that is we take standard deviation of $\log_{10}(F(w_k) - F(w_*))$. We then plot the curve $\pm$ this standard deviation. Details about all experimentation settings are gathered in Table S1.

Convergence. Figure 4a presents the convergence of each algorithm w.r.t. the number of iterations k. During first iterations all algorithms make fast progress.



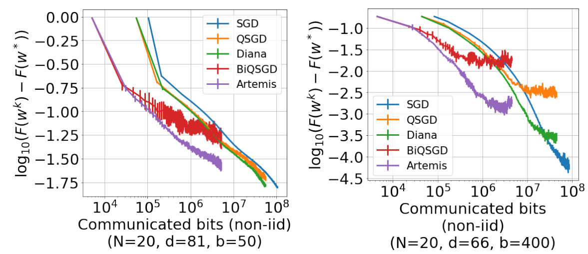
(a) LSR: $\sigma_*^2 \neq 0$ (b) LSR: X-axis in # bits.

Figure 4: Illustration of `Artemis` compared to existing algorithms on i.i.d. data. Best seen in colors.



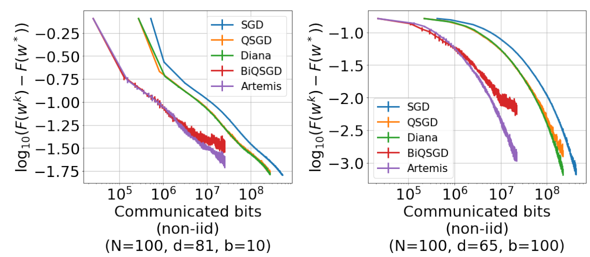
(a) LSR (i.i.d.): $\sigma_*^2 = 0$ (b) LR (non-i.i.d.): $\sigma_*^2 = 0$

Figure 5: Left: illustration of the linear convergence of all algorithms when $\sigma_*^2 = 0$ but $\sigma_{\mathrm{unif}}^2 > 0$, right: illustration of the memory benefits when $\sigma_* = 0$ but data is non-i.i.d. Best seen in colors.



(a) Superconduct (b) Quantum

Figure 6: **Real dataset (non-i.i.d.)**: $\sigma_* \neq 0$, $N = 20$ workers, $b > 1$ (1000 iter.). X-axis in # bits.



(a) Superconduct (b) Quantum

Figure 7: **Devices partial participation**: $\sigma_* \neq 0$, $N = 100$ devices and $p = 0.5$ (1000 iter.). X-axis in # bits.

However 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.

Complexity. On Figure 4b, the loss is plotted w.r.t. the theoretical number of bits exchanged after k iterations. This confirms that double compression should be the method of choice to achieve a reasonable precision (*w.r.t.* σ_*): for 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 Figures 4a and 5a. Saturation is observed in Figure 4a, 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. [2020] 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 4a with Figure 5a sheds light on the fact that the saturation level (before which double compression is indeed beneficial, as observed in Figure 4b) 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 S9 in Appendix C.1.1).

Non-i.i.d distributions and real dataset. While in Figures 4a, 4b and 5a, 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 5b, 6a and 6b we use **non-i.i.d. data** distributed in an unbalanced manner. None of the previous papers on compression directly illustrated the impact of non-i.i.d.-ness on simple examples, neither compare it with i.i.d. situations.

First on Figure 5b, we use the toy LR model, with $\sigma_* = 0$, observing linear convergence of `Artemis` and `Diana` and SGD (with increasing communication costs). We then consider the most general framework, with non i.i.d., $\sigma^* > 0$, higher-dimensional real data, and mini-batch b on Figures 6a and 6b, as well as Figure 7 which illustrates the situation where only a subsample of devices are active at each round. As predicted by Theorem 1, `Artemis` outperforms other algorithms.

We also perform experiments with *optimized* learning rates, see Figure S17.

5 Conclusion

We propose `Artemis`, an algorithm using a bidirectional compression to reduce the number of bits needed to perform FL. On top of compression, `Artemis` is enhanced by a memory mechanism which improves convergence over non-i.i.d. data. As devices partial participation is a classical settings of Federated Learning, our algorithm handles this use-case. 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 bounds by studying convergence in distribution of our algorithm. Altogether, this improves the understanding of compression and sheds light on challenges ahead.

More generally, this work is aligned with a global effort to make the usage of large scale Federated Learning sustainable by minimizing its environmental impact.

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).

References

N. Agarwal, A. T. Suresh, F. Yu, S. Kumar, and H. B. Mcmahan. cpSGD: Communication-efficient and differentially-private distributed SGD. *arXiv:1805.10559 [cs, stat]*, May 2018.

A. F. Aji and K. Heafield. Sparse Communication for Distributed Gradient Descent. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 440–445, 2017.

D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic. QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding. *arXiv:1610.02132 [cs]*, Dec. 2017.

D. Alistarh, T. Hoefler, M. Johansson, S. Khirirat, N. Konstantinov, and C. Renggli. The Convergence of Sparsified Gradient Methods. *arXiv:1809.10505 [cs, stat]*, Sept. 2018.

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

L. Bottou. On-line learning and stochastic approximations. 1999.

R. Caruana, T. Joachims, and L. Backstrom. KDD-cup 2004: results and analysis. 6(2):95–108. ISSN 1931-0145.

A. Dieuleveut, A. Durmus, and F. Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. *Annals of Statistics*, 2019.

H. Eichner, T. Koren, H. B. McMahan, N. Srebro, and K. Talwar. Semi-Cyclic Stochastic Gradient Descent. Apr. 2019.

P. Elias. Universal codeword sets and representations of the integers, Sept. 1975.

R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtarik. SGD: General Analysis and Improved Rates. *arXiv:1901.09401 [cs, math, stat]*, May 2019.

K. Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. 154:346–354. ISSN 0927-0256.

S. Horváth and P. Richtárik. A better alternative to error feedback for communication-efficient distributed learning. 2020.

S. Horváth, D. Kovalev, K. Mishchenko, S. Stich, and P. Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. 2019.

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

S. P. Karimireddy, Q. Rebjock, S. U. Stich, and M. Jaggi. Error Feedback Fixes SignSGD and other Gradient Compression Schemes. *arXiv:1901.09847 [cs, math, stat]*, May 2019.

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

J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. *arXiv:1610.02527 [cs]*, Oct. 2016.

L. Lannelongue, J. Grealey, and M. Inouye. Green algorithms: Quantifying the carbon emissions of computation. 2020.

T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated Optimization in Heterogeneous Networks. *arXiv:1812.06127 [cs, stat]*, Sept. 2019.

Z. Li, D. Kovalev, X. Qian, and P. Richtárik. Acceleration for Compressed Gradient Descent in Distributed and Federated Optimization. *arXiv:2002.11364 [cs, math]*, Feb. 2020.

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

L. v. d. Maaten and G. Hinton. Visualizing data using t-SNE. 9:2579–2605. ISSN ISSN 1533-7928.

H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. Feb. 2017.

S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009. ISBN 0521731828, 9780521731829.

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

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer US, 2004. ISBN 978-1-4020-7553-7.

K. K. Patel and A. Dieuleveut. Communication tradeoffs for synchronized distributed SGD with large step size. *arXiv:1904.11325 [cs, math, stat]*, Apr. 2019.

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

A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani. FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization. In *International Conference on Artificial Intelligence and Statistics*, pages 2021–2031. PMLR, 2019.

Report. Speedtest global index – monthly comparisons of internet speeds from around the world, anonymous authors, 2020.

H. Robbins and S. Monro. A Stochastic Approximation Method. *Annals of Mathematical Statistics*, 22(3): 400–407, Sept. 1951. ISSN 0003-4851, 2168-8990.

F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. pages 1058–1062, Jan. 2014.

H. Tang, X. Lian, C. Yu, T. Zhang, and J. Liu. DoubleSqueeze: Parallel Stochastic Gradient Descent with Double-Pass Error-Compensated Compression. *arXiv:1905.05957 [cs]*, June 2019.

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

W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li. TernGrad: Ternary Gradients to Reduce Communication in Distributed Deep Learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1509–1519. Curran Associates, Inc., 2017.

J. Wu, W. Huang, J. Huang, and T. Zhang. Error Compensated Quantized SGD and its Applications to Large-scale Distributed Optimization. *arXiv:1806.08054 [cs]*, June 2018.

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

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. 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.

S. Zheng, Z. Huang, and J. Kwok. Communication-Efficient Distributed Blockwise Momentum SGD with Error-Feedback. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 11450–11460. Curran Associates, Inc., 2019.

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

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

`Artemis`: tight convergence guarantees for bidirectional compression in Federated Learning

Supplementary material

In this appendix, we provide additional details to our work. In Appendix A, we give more details on devices partial participation, we describe the s-quantization scheme used in our experiments and we define the filtrations used in following 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

A Additional details about the `Artemis` framework

The aim of this section is threefold. First, we details the choice made by `Artemis` to handle the case of devices partial participation. Secondly, we provide supplementary details about the quantization scheme used in our work. We also explain (based on Alistarh et al. [2017]) how quantization combined with `Elias` code [see Elias, 1975] reduces the required number of bits to send information. Finally, we define the filtrations used in our proofs and give their resulting properties.

A.1 Device sampling

In this subsection, we detail how to handle in practice devices partial participation. Namely, we describe the various possible tactics to tackle this issue and we explain the "catching-up" process used in `Artemis` that allows to synchronize the models.

Must of the algorithms handling with device sampling (Reisizadeh et al. [2019]; Eichner et al. [2019]; Li et al. [2019]; McMahan et al. [2017]) send the model. This is an easy setting because it does not require to synchronize the local models with the remote model. However, `Artemis` only sends the gradient and not the model from the central server to the remote workers. This configuration is harder because the local model is updated using the values sent. As a consequence, handling device sampling is not straightforward. Indeed if a given worker

$i \in [\![1, N]\!]$ is switched off for few iterations, it will receive at iteration $k \in \mathbb{N}$ a value Ω_{k+1,S_k} (compression of the gradient) that corresponds to a model different from the one it owns. And thus the local update will be meaningless.

Thus, there is a need for model synchronization. Below, we list the few possible tactics, their advantages and their drawbacks:

- Send back Ω_{k+1,S_k} solely to devices that have participated to the round. This leads to unsynchronized models on each device (i.e all local models are different and disconnected from the received updates), thus the local updates are meaningless and the learning can not converge.
- Send back Ω_{k+1,S_k} to all devices regardless to which ones are participating or not. This case would be possible only if all devices were always switched on (and could always receive broadcast) but were not always participating to training. This use-case can not be applied to a smartphone network where devices can be switched off, disconnected from the internet, or just busy.
- Keep in memory both the sequence of $(\Omega_{k+1,S_k})_{k\in\mathbb{N}}$, and the indices $(k_i)_{i\in[\![1,N]\!]}$ of the last update for each device. When a device is switched off we stop to communicate the updates, but as soon at it is available again, we send the sequence of $(\Omega_{k+1,S_k})_{k\in\mathbb{N}}$ it missed. This approach has two main drawback. 1) The sequence of $(\Omega_{k+1,S_k})_{k\in\mathbb{N}}$ and the index of last update for each device must be stored on the central server leading to a memory usage increase. 2) If the sequence of $(\Omega_{k+1,S_k})_{k\in\mathbb{N}}$ is too big, it may cheaper to send the complete last model instead. The asset of this method is that it is easily achievable in the real world. This approach is the one used for `Artemis`. To implement this solution, there are two ways:
  1. Keep all $(h^i)_{i\in[\![1,N]\!]}$ on both central server and remote nodes. When all the information are aggregated, we use only the memories related to the active nodes. This is the most natural way, and this is how `Artemis` works.
  2. Don't keep all $(h^i)_{i\in[\![1,N]\!]}$ on the central server, keep only a vector h. This approach is somehow closer to the way `Diana` has been designed by Mishchenko et al. [2019]. In this work, they define for k in $\mathbb{N}$, $h_{k+1} = h_k + \alpha\widehat{\Delta}_k$. When all devices are participating this is equivalent to the first option. But, as soon as some devices are missing to the training step, for k in $\mathbb{N}$ the vector h_k becomes unsynchronized with the local memories $(h_k^i)_{i\in[\![1,N]\!]}$. However, far from degrading performance, it leads to introduce a moment in the algorithm that considerably speed-up the convergence. Indeed, when all the information is aggregated on central server and combined with the central memory h_k, workers that didn't participate for a few rounds are taking benefit of the past updates they missed. This phenomenon has been observed in experiments, but while its helps the convergence of the algorithm, it is also harder to analyse theoretically.
- The last option would be to send the full last model when a device is switched on. This way, it carries out a correct computation and can participate to the learning. But this approach short-circuits the bidirectional compression. For instance, if devices are participating at one round on two, this will be as if there is a one-way compression.

A.2 Quantization scheme

In the following, we define the s-quantization operator $\mathcal{C}_s$ which we use in our experiments. After giving its definition, we explain [based on Alistarh et al., 2017] how it helps to reduce the number of bits to broadcast.

Definition 1 (s-quantization operator). *Given $\Delta \in \mathbb{R}^d$, the s-quantization operator $\mathcal{C}_s$ is defined by:*

$$\mathcal{C}_s(\Delta) := sign(\Delta) \times \|\Delta\|_2 \times \frac{\psi}{s} .$$

$\psi \in \mathbb{R}^d$ is a random vector with j-th element defined as:

$$\psi_j := \begin{cases} l+1 & \text{with probability } s\frac{|\Delta_j|}{\|\Delta\|_2} - l \\ l & \text{otherwise} . \end{cases}$$

where the level l is such that $\dfrac{\Delta_i}{\|\Delta\|_2} \in \left[\dfrac{l}{s}, \dfrac{l+1}{s}\right]$.

The s-quantization scheme verifies Assumption 5 with $\omega_{\mathcal{C}} = \min(d/s^2, \sqrt{d}/s)$. Proof can be found in [Alistarh et al., 2017, see Appendix A.1].

Now, for any vector $v \in \mathbb{R}^d$, we are in possession of the tuple $(\|v\|^2, \phi, \psi)$, where ϕ is the vector of signs of $(v_i)_{i=1}^d$, and ψ is the vector of integer values $(\psi_j)_{j=1}$. To broadcast the quantized value, we use the `Elias` encoding Elias [1975]. Using this encoding scheme, it can be shown (Theorem 3.2 of Alistarh et al. [2017]) that:

Proposition S1. *For any vector v, the number of bits needed to communicate $\mathcal{C}_s(v)$ is upper bounded by:*

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

The final goal of using memory for compression is to quantize vectors with $s = 1$. It means that we will employ $O(\sqrt{d}\log d)$ bits per iteration instead of $32d$, which reduces by a factor $\frac{\sqrt{d}}{\log d}$ 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 devices. It means that at each iteration, we need to communicate $2 \times N \times 32d$ bits if compression is not used. With a single compression process like in Mishchenko et al. [2019]; Li et al. [2020]; Wu et al. [2018]; Agarwal et al. [2018]; Alistarh et al. [2017], we need to broadcast

$$\begin{aligned} O\left(32Nd + N\sqrt{d}\log d\right) &= O\left(Nd\left(1 + \frac{\log d}{\sqrt{d}}\right)\right) \\ &= O\left(Nd\right)\,. \end{aligned}$$

But with a bidirectional compression, we only need to broadcast $O\left(2N\sqrt{d}\log d\right)$.

Time complexity analysis of simple vs double compression for the 1-quantization schema. Using quantization with $s = 1$, and then the `Elias` code [defined in Elias, 1975] to communicate between servers, leads to reduce from $O(Nd)$ to $O(N\sqrt{d}\log(d))$ the number of bits to send, for each direction. Getting an estimation of the total time complexity is difficult and inevitably dependant of the considered application. Indeed, as highlighted by Figure 1, download and upload speed are always different. The biggest measured difference between upload and download is found in Europa for mobile broadband ; their ratio is around 3.5.

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 $O\left(\frac{Nd}{v_d} + \frac{N\sqrt{d}\log(d)}{v_u}\right) \approx O\left(\frac{Nd}{\rho v_u}\right)$ seconds, while for a bidirectional one it takes only $O\left(\frac{N\sqrt{d}\log(d)}{v_d} + \frac{N\sqrt{d}\log(d)}{v_u}\right) \approx O\left(\frac{N\sqrt{d}\log(d)}{v_u}\right)$ seconds.

In other words, unless ρ is really large (which is not the case in practice as stressed by Figure 1, double compression reduces by several orders of magnitude the global time complexity, and bidirectional compression is by far superior to unidirectional.

A.3 Filtrations

In this section we provide some explanations about filtrations - especially a rigorous definition - and how it is used in the proofs of Theorems 2, 3, S4 and S5. We recall that we denoted by $\omega_{\mathcal{C}}^{\mathrm{up}}$ and $\omega_{\mathcal{C}}^{\mathrm{dwn}}$ the variance factors for respectively uplink and downlink compression.

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

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

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

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

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

Concerning randomness in our algorithm, it comes from four sources:

$$w_k \xrightarrow{\xi_{k+1}^i} g_{k+1}^i \xrightarrow{\epsilon_{k+1}^i} \widehat{g}_{k+1}^i \xrightarrow{B_k^i} \widehat{g}_{k+1,S_k} = \sum_{i \in S_k} \widehat{g}_{k+1}^i \xrightarrow{\epsilon_{k+1}} G_{k+1,S_k} = \mathcal{C}\left(\widehat{g}_{k+1,S_k}\right)$$

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

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

$$\forall k \in \mathbb{N}\,,\ \forall i \in [\![0, ..., N]\!], \quad g_{k+1}^i = \nabla F_i(w_k) + \xi_{k+1}^i(w_k)\,, \text{ with } \mathbb{V}(\xi_{k+1}^i) \text{ bounded.}$$

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 the associated noise is ϵ_{k+1}^i with $\mathbb{V}(\epsilon_{k+1}^i(\Delta_k^i)) \leq \omega_{\mathcal{C}}^{\mathrm{up}} \left\|\Delta_k^i\right\|^2$, where $\omega_{\mathcal{C}}^{\mathrm{up}} \in \mathbb{R}^*$ is defined by the uplink compression schema (see Assumption 5). And it follows that:

$$\forall k \in \mathbb{N},\ \forall i \in [\![0, ..., N]\!], \quad \widehat{\Delta}_k^i = \Delta_k^i + \epsilon_k^i(\Delta_k^i) \Longleftrightarrow \widehat{g}_{k+1}^i = g_{k+1}^i + \epsilon_{k+1}^i(\Delta_k^i)\,.$$

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 $\widehat{g}_{k+1,S_k} \in \mathbb{R}^d$, then the associated noise is $\epsilon_{k+1}(\widehat{g}_{k+1,S_k})$ with $\mathbb{V}(\epsilon_{k+1}) \leq \omega_{\mathcal{C}}^{\mathrm{dwn}} \left\|\widehat{g}_{k+1,S_k}\right\|^2$. There is:

$$\forall k \in \mathbb{N}, \quad G_{k+1,S_k} = \mathcal{C}_s(\widehat{g}_{k+1,S_k}) = \widehat{g}_{k+1,S_k} + \epsilon_{k+1}((\widehat{g}_{k+1,S_k}))\,.$$

4. Random sampling. This randomness corresponds to the partial participation of each device. We recall that according to Assumption 6, each device has a probability p of being active. For k in $\mathbb{N}$, for i in $[\![1, N]\!]$, we note $B_k^i \sim \mathcal{B}(p)$ the Bernoulli random variable that marks if a device is active or not at step k.

This "succession of noises" in the algorithm is illustrated in Figure S1. In order to handle these four sources of randomness, we define five 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, $(\mathcal{B}_k)_{k\in\mathbb{N}}$ the filtration associated to the random device participation randomness. For $k \in \mathbb{N}^*$, we define:*

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

with

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

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

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

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

  However, $(\mathcal{B}_k)_{k\in\mathbb{N}}$ is independent of the other filtrations.

- $\mathcal{I}_k = \sigma\left(\mathcal{H}_k \cup \mathcal{B}_{k-1}\right) = \sigma(\Gamma_k)$, and the aim is to express the expectation w.r.t. all randomness i.e $\mathcal{I}_k$.
- w_k is $\mathcal{I}_k$-measurable.

- $g_{k+1}(w_k)$ is $\mathcal{F}_{k+1}$-measurable.
- $\widehat{g}_{k+1}(w_k)$ is $\mathcal{G}_{k+1}$-measurable.
- B^i_{k-1} is $\mathcal{B}_{k-1}$-measurable.
- g_{k+1,S_k} , $\widehat{g}_{k+1,S_k}$ and G_{k+1,S_k} are respectively $\sigma(\mathcal{F}_{k+1} \cup \mathcal{B}_k)$-measurable, $\sigma(\mathcal{G}_{k+1} \cup \mathcal{B}_k)$-measurable and $\sigma(\mathcal{H}_{k+1} \cup \mathcal{B}_k)$-measurable. Note that $\mathcal{F}_{k+1}$ contains Γ_k, and thus all $(B^i_{k-1})_{i=1}^N$, but does not contain all the $(B^i_k)_{i=1}^N$.

As a consequence, we have Propositions S2 to S8. Please, take notice that for sake of clarity Propositions S2 to S6 are stated without taking into account the random participation S_k. All this proposition remains identical when adding partial participation, the results only have to be expressed w.r.t. active nodes S_k.

Below Proposition S2 gives the expectation over stochastic gradients conditionally to σ-algebras $\mathcal{H}_k$ and $\mathcal{F}_{k+1}$.

Proposition S2 (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.):*

$$\left\{ \begin{array}{ll} \mathbb{E}\left[g^i_{k+1} \;\middle|\; \mathcal{F}_{k+1}\right] & = g^i_{k+1} \\ \mathbb{E}\left[g^i_{k+1} \;\middle|\; \mathcal{H}_k\right] & = \nabla F_i(w_k)\,, \end{array} \right.$$

which leads to:

$$\left\{ \begin{array}{ll} \mathbb{E}\left[g_{k+1} \mid \mathcal{F}_{k+1}\right] & = g_{k+1} \\ \mathbb{E}\left[g_{k+1} \mid \mathcal{H}_k\right] & = \nabla F(w_k)\,. \end{array} \right.$$

Proposition S3 gives expectation of uplink compression (information sent from remote devices to central server) conditionally to σ-algebras $\mathcal{F}_{k+1}$ and $\mathcal{G}_{k+1}$.

Proposition S3 (Uplink Compression Expectation)**.** *Let $k \in \mathbb{N}$ and $i \in [\![1, N]\!]$. Recall that $\widehat{g}^i_k = g^i_k + \epsilon^i_k$, then on each local device $i \in [\![1, N]\!]$, we have a.s:*

$$\left\{ \begin{array}{l} \mathbb{E}\left[\widehat{g}^i_{k+1} \;\middle|\; \mathcal{G}_{k+1}\right] = \widehat{g}^i_{k+1} \\ \mathbb{E}\left[\widehat{g}^i_{k+1} \;\middle|\; \mathcal{F}_{k+1}\right] = g^i_{k+1}\,, \end{array} \right.$$

which leads to

$$\left\{ \begin{array}{l} \mathbb{E}\left[\widehat{g}_{k+1} \mid \mathcal{G}_{k+1}\right] = \widehat{g}_{k+1} \\ \mathbb{E}\left[\widehat{g}_{k+1} \mid \mathcal{F}_{k+1}\right] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \widehat{g}^i_{k+1} \;\middle|\; \mathcal{F}_{k+1}\right] = g_{k+1}\,. \end{array} \right.$$

From Assumption 5, if follows that variance over uplink compression can be bounded as expressed in Proposition S4.

Proposition S4 (Uplink Compression Variance)**.** *Let $k \in \mathbb{N}$ and $i \in [\![1, N]\!]$. Recall that $\Delta^i_k = g^i_{k+1} + h^i_k$, using Assumption 5 following hold a.s:*

$$\mathbb{E}\left[\|\widehat{\Delta}^i_{k+1} - \Delta^i_{k+1}\|^2 \;\middle|\; \mathcal{F}_{k+1}\right] \leq \omega_{\mathcal{C}}^{\mathrm{up}} \|\Delta^i_{k+1}\|^2 \tag{S1}$$

$$(\Longleftrightarrow \qquad \mathbb{E}\left[\|\widehat{g}^i_{k+1} - g^i_{k+1}\|^2 \mid \mathcal{F}_{k+1}\right] \leq \omega_{\mathcal{C}}^{\mathrm{up}} \|g^i_{k+1}\|^2 \text{ when no memory })\,. \tag{S2}$$

Concerning downlink compression (information sent from central server to each node), Proposition S5 gives its expectation w.r.t σ-algebras $\mathcal{G}_{k+1}$ and $\mathcal{H}_{k+1}$.

Proposition S5 (Downlink Compression Expectation)**.** *Let $k \in \mathbb{N}$, recall that $G_{k+1} = \mathcal{C}_{\mathrm{dwn}}(\widehat{g}_{k+1}) = \widehat{g}_{k+1} + \epsilon_k$, then a.s:*

$$\left\{ \begin{array}{l} \mathbb{E}\left[G_{k+1} \mid \mathcal{H}_{k+1}\right] = G_{k+1} \\ \mathbb{E}\left[G_{k+1} \mid \mathcal{G}_{k+1}\right] = \widehat{g}_{k+1}\,. \end{array} \right.$$

Next proposition states that downlink compression can be bounded as for Proposition S4.

Proposition S6 (Downlink Compression Variance)**.** *Let $k \in \mathbb{N}$, using Assumption 5 following holds a.s:*

$$\mathbb{E}\left[\|G_{k+1} - \widehat{g}_{k+1}\|^2 \mid \mathcal{G}_{k+1}\right] \leq \omega_{\mathcal{C}}^{\mathrm{dwn}} \|\widehat{g}_{k+1}\|^2\,.$$

Now, we give in Propositions S7 and S8 the expectation and the variance w.r.t. devices random sampling noise.

Proposition S7 (Expectation of device sampling). *Let $k \in \mathbb{N}$, let's note $a_{k+1} = \frac{1}{N}\sum_{i=1}^N a_{k+1}^i$ and $a_{k+1,S_k} = \frac{1}{pN}\sum_{i \in S_k} a_{k+1}^i$, where $(a_{k+1}^i)_{i=0}^N \in (\mathbb{R}^d)^N$ are N random variables independent of each other and $\mathcal{J}_{k+1}$-measurable, for a σ-field $\mathcal{J}_{k+1}$ s.t. $(B_k^i)_{i=1}^N$ are independent of $\mathcal{J}_{k+1}$. We have a.s:*

$$\mathbb{E}\left[a_{k+1,S_k} \mid \mathcal{J}_{k+1}\right] = a_{k+1} \,.$$

The vector a_{k+1} (resp. the σ-field $\mathcal{J}_{k+1}$) may represent various objects, for instance : g_{k+1}, $\widehat{g}_{k+1}$, G_{k+1} (resp. $\mathcal{F}_{k+1}$, $\mathcal{G}_{k+1}$, $\mathcal{H}_{k+1}$).

Proof. For any $k \in \mathbb{N}^*$, we have that:

$$\begin{aligned}
\mathbb{E}\left[a_{k+1,S_k} \mid \mathcal{J}_{k+1}\right] &= \mathbb{E}\left[\frac{1}{pN}\sum_{i\in S_k} a_{k+1}^i \,\middle|\, \mathcal{J}_{k+1}\right] = \mathbb{E}\left[\frac{1}{pN}\sum_{i=0}^N a_{k+1}^i B_k^i \,\middle|\, \mathcal{J}_{k+1}\right] \\
&= \frac{1}{pN}\sum_{i=0}^N \mathbb{E}\left[a_{k+1}^i B_k^i \mid \mathcal{J}_{k+1}\right] \text{ by linearity of the expectation,} \\
&= \frac{1}{pN}\sum_{i=0}^N a_{k+1}^i \mathbb{E}\left[B_k^i \mid \mathcal{J}_{k+1}\right] \text{because } (a_{k+1}^i)_{i=1}^N \text{ are } \mathcal{J}_{k+1}\text{-measurable,} \\
&= \frac{1}{N}\sum_{i=0}^N a_{k+1}^i = a_{k+1} \text{ because } (B_k^i)_{i=1}^N \text{ are independent of } \mathcal{J}_{k+1},
\end{aligned}$$

which allows to conclude. □

Proposition S8 (Variance of device sampling). *Let $k \in \mathbb{N}^*$, with the same notation as Proposition S7, we have a.s:*

$$\mathbb{V}\left[a_{k+1,S_k} \mid \mathcal{J}_{k+1}\right] = \frac{1-p}{pN^2}\sum_{i=0}^N \left\|a_{k+1}^i\right\|^2 .$$

Proof. Let $k \in \mathbb{N}^*$,

$$\begin{aligned}
\mathbb{V}\left[a_{k+1,S_k} \mid \mathcal{J}_{k+1}\right] &= \mathbb{V}\left[\frac{1}{pN}\sum_{i=0}^N a_{k+1}^i B_k^i \,\middle|\, \mathcal{J}_{k+1}\right] \\
&= \frac{1}{p^2N^2}\sum_{i=0}^N \mathbb{V}\left[a_{k+1}^i B_k^i \mid \mathcal{J}_{k+1}\right] \text{because } (B_k^i)_{i=1}^N \text{ are independent,} \\
&= \frac{1}{p^2N^2}\sum_{i=0}^N \left\|a_{k+1}^i\right\|^2 \mathbb{V}\left[B_k^i \mid \mathcal{J}_{k+1}\right] \text{because } (a_{k+1}^i)_{i=1}^N \text{ are } \mathcal{J}_{k+1}\text{-measurable,} \\
&= \frac{1-p}{pN^2}\sum_{i=0}^N \left\|a_{k+1}^i\right\|^2 \text{ because } (B_k^i)_{i=1}^N \text{ are independent of } \mathcal{J}_{k+1}.
\end{aligned}$$

□

B Bandwidth speed

The dataset is pickled from a study carried out by *Speedtest.net* [see Report, 2020]. 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 to S5.

In Figures S3 to S5, unlike Figure S2 which is the same figure as the one in Section 1, we do not aggregate data by countries of a same continents. This allows to analyse the speeds ratio between upload and download with the

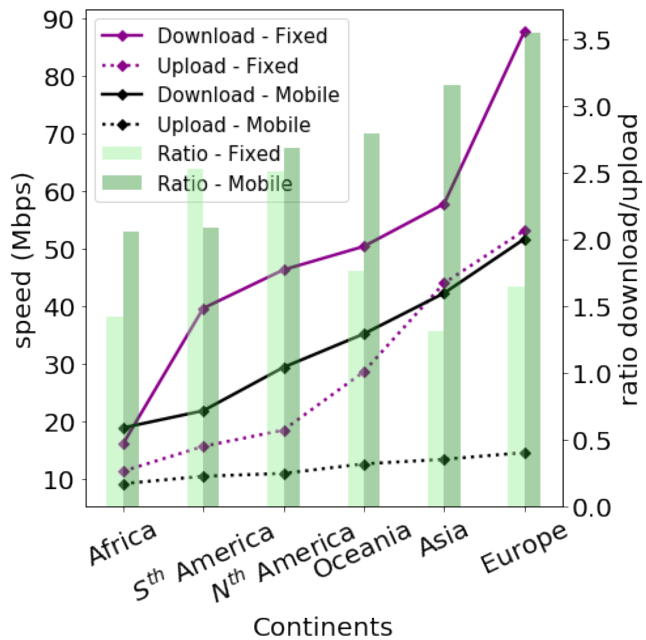


Figure S2: Duplicate of Figure 1. Left axis: upload and download speed for mobile and fixed broadband. Left axis: speeds (in Mbps), right axis: ratio (green bars). The dataset is gathered from *Speedtest.net*, see Report [2020].

proper value of each countries. Looking at Figures S3 to S5, 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 1 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.

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* [see Hamidieh, 21263 points, 81 features] and *quantum* [see Caruana et al., 50,000 points, 65 features]. The aim of using synthetic datasets is mainly to underline

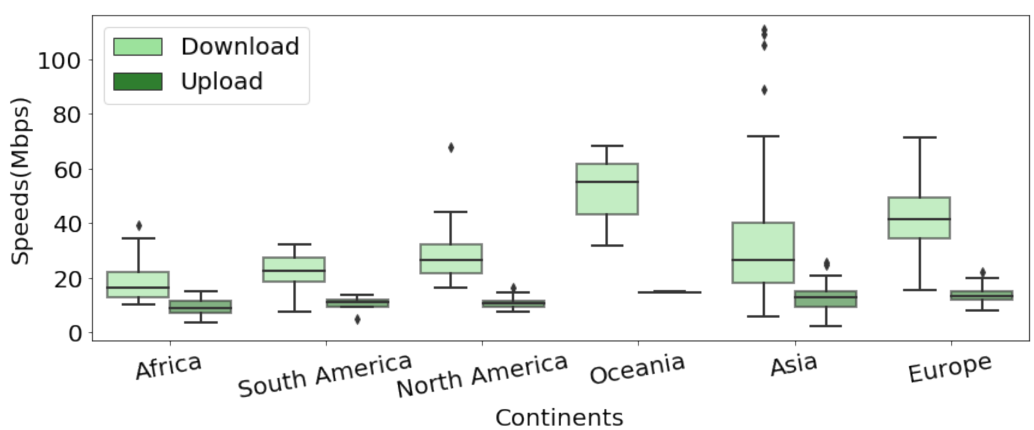


Figure S3: Upload/download speed (in Mbps) for mobile broadband. Best seen in colors.

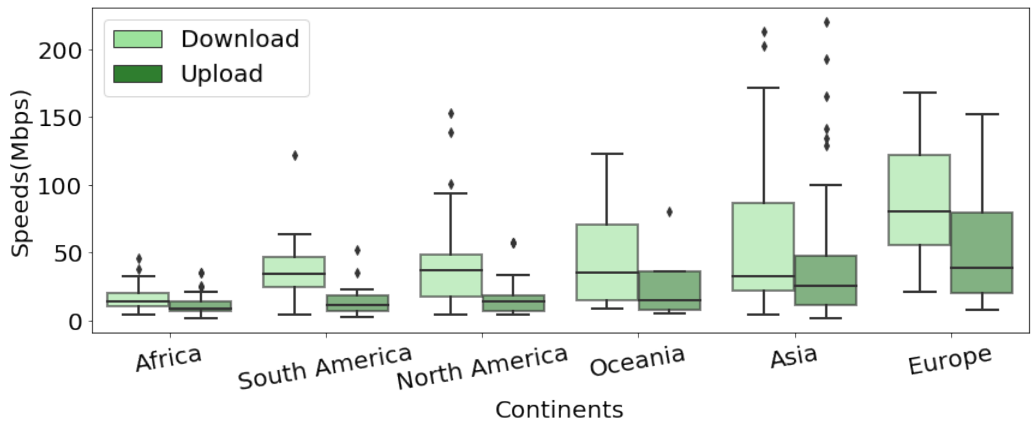


Figure S4: Upload/download speed (in Mbps) for fixed broadband. Best seen in colors.

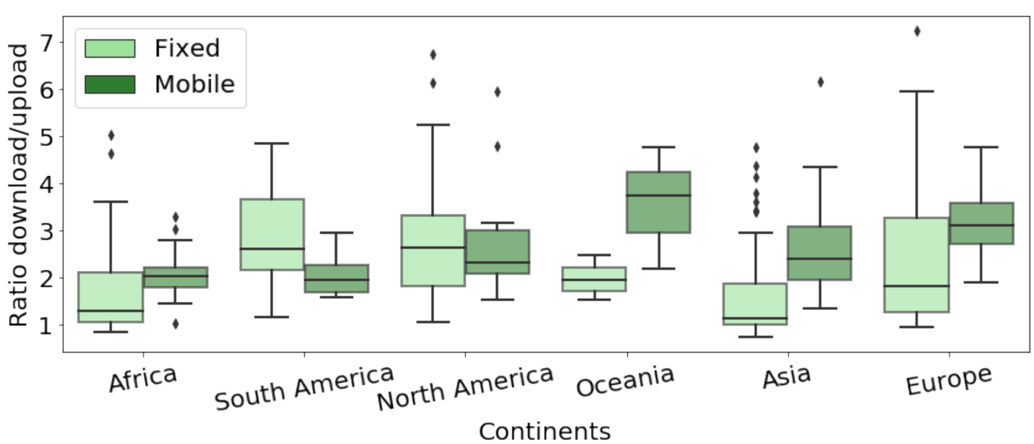
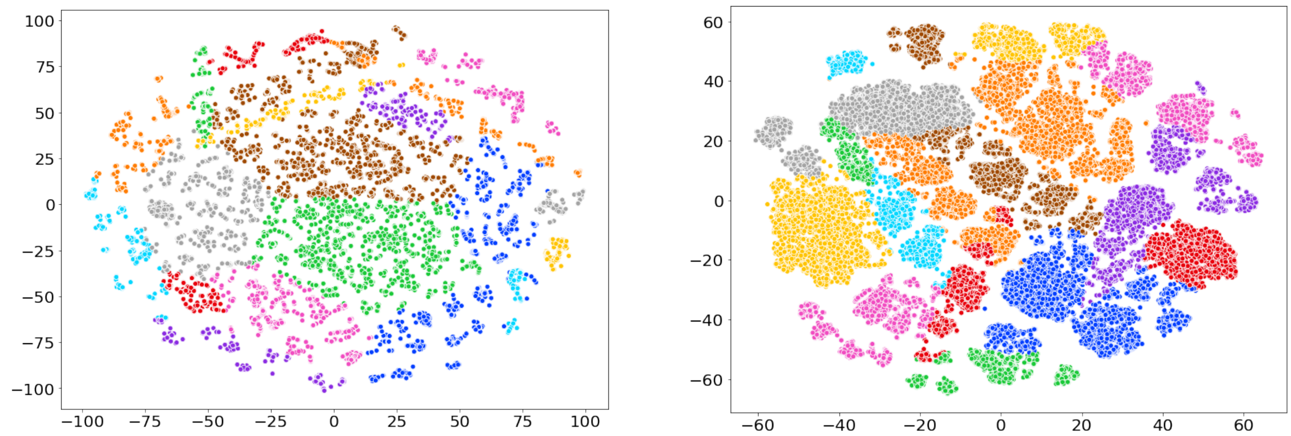


Figure S5: Distribution of the download/upload speeds ratio by continents. Best seen in colors.



(a) Superconduct dataset: 20 cluster. Each cluster has between 250 and 3900 points with a median at 750 points.

(b) Quantum dataset: 20 clusters. Each cluster has between 900 and 10500 points with a median at 2300 points.

Figure S6: TSNE representations. Best seen in colors.

the properties resulting from Theorems 1 to 3.

With the real datasets, in order to simulate non-i.i.d. data and to make the experiments closer to real-life usage, we split the dataset in heterogeneous groups using a Gaussian mixture clustering on TSNE representations (defined by Maaten and Hinton). Thus, the data is highly non-i.i.d. *and* unbalanced over devices. We plot on Figure S6 the TSNE representation of the two real datasets.

We use the same 1-quantization scheme (defined in Appendix A.2, $s = 1$ is the most drastic compression) for both uplink and downlink, and thus, we consider that $\omega_{\mathcal{C}}^{\mathrm{up}} = \omega_{\mathcal{C}}^{\mathrm{dwn}}$. 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 $\log_{10}(F(w_k) - F(w_*))$, with k in $\mathbb{N}$. All experiments have been run 5 times and averaged before displaying the curves. We plot error bars on all figures. To compute error bars we use the standard deviation of the logarithmic difference between the loss function at iteration k and the objective loss, that is we take standard deviation of $\log_{10}(F(w_k) - F(w_*))$. We then plot the curve $\pm$ this standard deviation.

In Table S1 we summarize all hyperparameters used to generate experimental figures presented in this paper. All the code is available on this repository: `https://github.com/philipco/artemis-bidirectional-compression`.

C.1 Synthetic dataset

We build two different synthetic dataset 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. As explained in Section 1, 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 $N = 10$ 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.

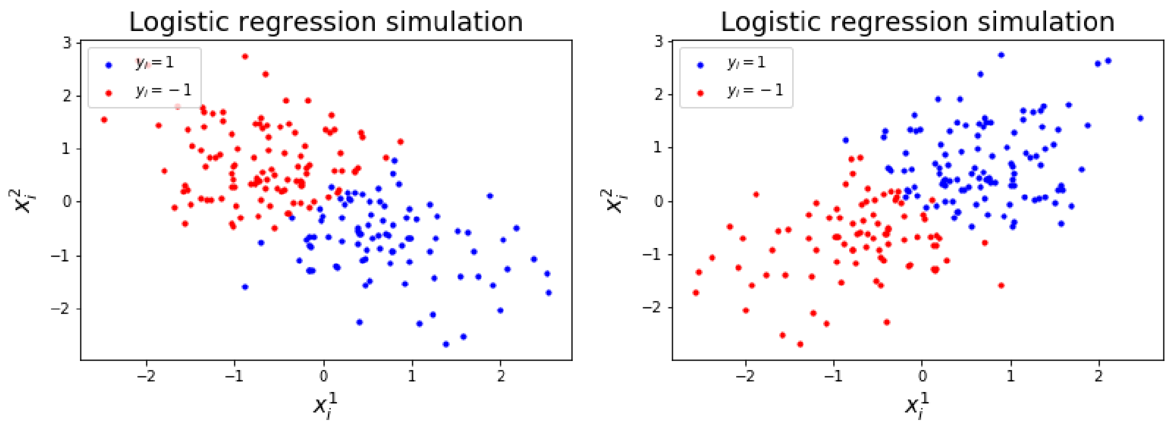
Choice of the step size for synthetic dataset. For stochastic descent, we use a step size $\gamma = \frac{1}{L\sqrt{k}}$ with k the number of iteration, and for the batch descent we choose $\gamma = \frac{1}{L}$.

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 then, for all j in $[\![1, n_i]\!]$, we have: $y_j^i = \left\langle w \mid x_j^i \right\rangle + e_i$ 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

Figure	n_i	N	d	K	σ_*^2	Task	i.i.d	p	Av.	b	Notebook	Illustrates
Figures 4a, 4b and S8	200	20	20	100	$\neq 0$	LSR	y	1	✗	1	with_noise	I.i.d data and saturation at high level.
Figures 5a and S9	200	20	20	100	0	LSR	y	1	✗	1	without_noise	Linear convergence of `Artemis` in absence of noise at the optimum
Figures 5b and S10	200	10	2	400	0	LR	n	1	✗	full batch	logistic_ deterministic	Benefits of memory over non-i.i.d data
Figure S11	200	20	2	400	0	LR	n	1	✓	full batch	logistic_ deterministic	Benefits of memory over non-i.i.d data with averaging
Figures 6a and S12b	var	20	81	1000	$\neq 0$	LSR	n	1	✗	50	superconduct	Benefits of memory over non-i.i.d data with real-life data
Figure S12a	var	20	81	100	$\neq 0$	LSR	n	1	✗	1	superconduct	Real dataset with $b = 1$
Figures 7a and S14a	var	100	81	1000	$\neq 0$	LSR	n	0.5	✗	10	s.-devices-sampling	Device partial participation
Figures 6b and S13b	var	20	65	1000	$\neq 0$	LR	n	1	✗	400	quantum	Benefits of memory over non-i.i.d data with real-life data
Figure S13a	var	20	65	100	$\neq 0$	LR	n	1	✗	1	quantum	Real dataset with $b = 1$
Figures 7b and S14b	var	100	65	1000	$\neq 0$	LR	n	0.5	✗	100	q.-devices-sampling	Device partial participation
Figure S15	var	20	81/65	250	$\neq 0$	LSR/LR	n	1	✗	50/400	s./q.-gamma_limit	Loss by step size for each algorithm
Figure S16	var	20	81/65	250	$\neq 0$	LSR/LR	n	1	✗	50/400	s./q.-gamma_limit	Loss of `Artemis` after 250 iter. for different step size
Figure S17	var	20	81/65	250	$\neq 0$	LSR/LR	n	1	✗	50/400	s./q.-gamma_limit	Loss with optimal step size

Table S1: List of hyperparameters used to generate the experimental figures, we hightlight in green the main difference of the figure compared to other experiments. All figures has been obtained after 5 runs, the average is plotted with the standard deviation. n_i is the number of points on each worker (var means the the number of points is variable), N is the number of machines, d is the dimension, K is the number of iterations, σ^2 is the variance at the optimum, LSR = Least Squares regression, LR = Logistic regression. In the “notebook” column, we indicate in which notebook the figure has been generated.



(a) Dataset 1

(b) Dataset 2

Figure S7: Data distribution for logistic regression to simulate non-i.i.d. data. Half of the device hold first dataset, and the other half the second one.

device i in $[\![1, N]\!]$ and all j in $[\![1, n_i]\!]$:

$$g_j^i(w) - F_i(w) = \underbrace{(X_j^i X_j^{i^T} - \mathbb{E} X^i X^{i^T})(w - w_*)}_{\text{multiplicative noise equal to 0 in } w_*} + \underbrace{(X_j^{i^T} w_* - Y_j^i)}_{\sim \mathcal{N}(0, \lambda^2)} X_j^i \tag{S3}$$

This is why, if we set $\lambda = 0$ and evaluate eq. (S3) 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 Figures 5a and S9. Otherwise, we set $\lambda^2 = 0.4$.

For non-i.i.d., 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 $[\![1, n_i]\!]$, $y_k^i = \mathcal{R}\left(\mathrm{Sigm}\left(\left\langle w_{(i \mod 2)+1} \,\middle|\, x^k_{(i \mod 2)+1} \right\rangle\right)\right) \in \{-1, +1\}$. 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:$x \mapsto \frac{\mathrm{e}^x}{1 + \mathrm{e}^x}$. These two distributions are presented on Figure S7.

C.1.1 Least-square regression

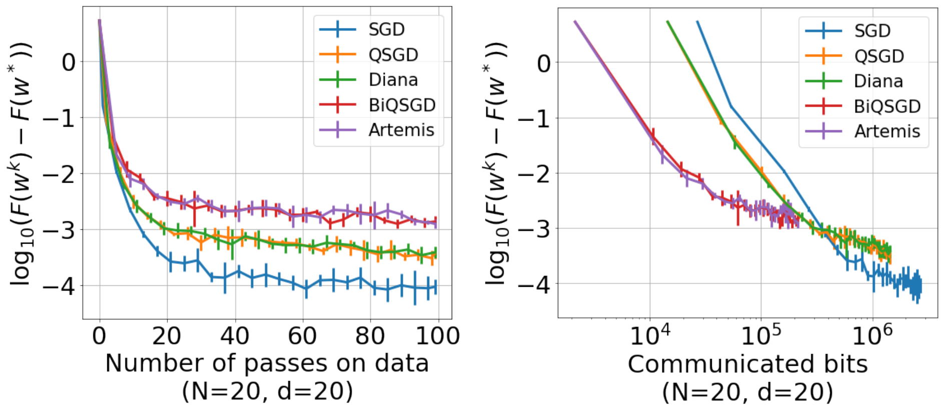
In this section, we present all figures generated using Least-Square regression. Figure S8 corresponds to Figures 4a and 4b, and Figure S9 corresponds to Figure 5a.

As explained in the main of the paper, in the case of $\sigma_* \neq 0$ (Figure S8), algorithm 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 S9), all algorithms (regardless of memory), converge at a linear rate exactly as classical SGD.

C.1.2 Logistic regression

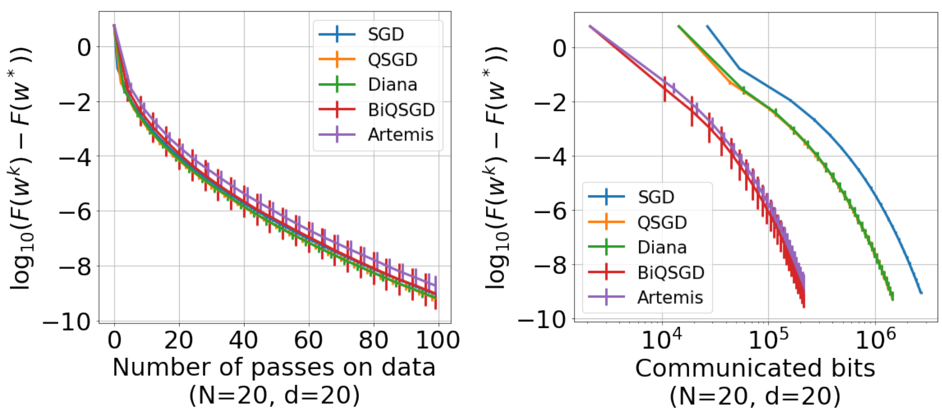
In this section, we present all figures generated using a logistic regression model. Figure S10 corresponds to Figure 5b. Data is non-i.d.d. and we use a full batch gradient descent to get $\sigma_* = 0$ to shed into light the impact of memory over convergence.

Figure S11 is using same data and configuration as Figure S10, except that *it is combined with 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.



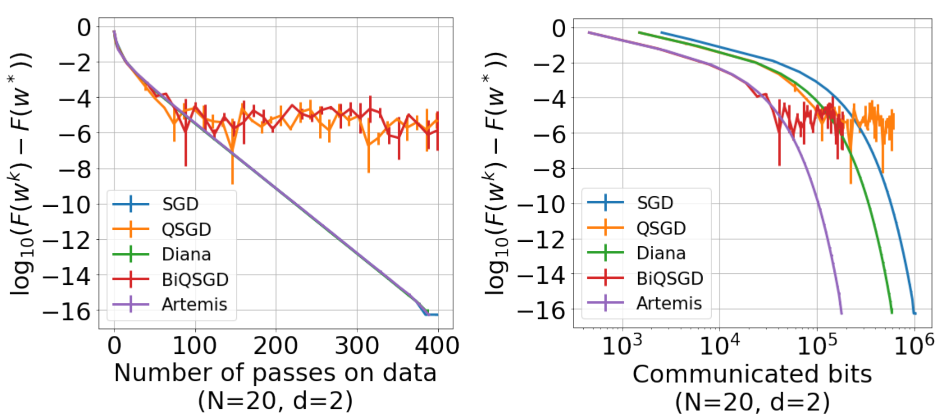
(a) LSR: $\sigma_*^2 \neq 0$ (b) X-axis in # bits.

Figure S8: **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.



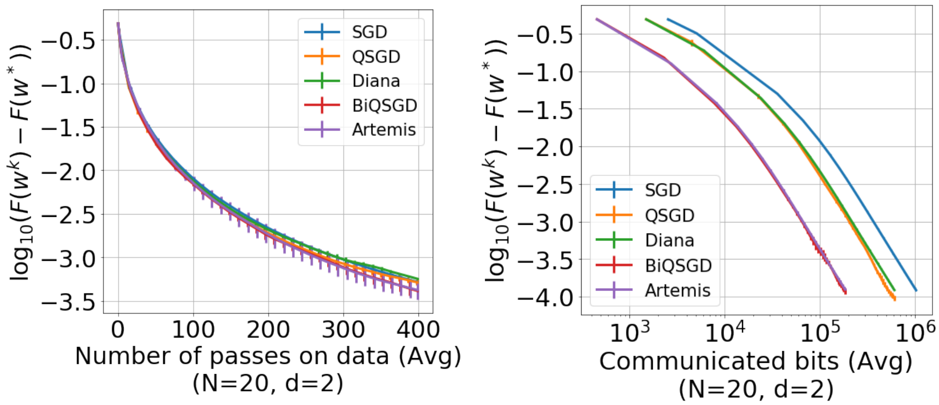
(a) LSR: $\sigma_*^2 = 0$ (b) X-axis in # bits.

Figure S9: **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 S10.



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

Figure S10: **Synthetic dataset, Logistic Regression on non-i.i.d. data** using a batch gradient descent (to get $\sigma_* = 0$). The benefit of memory is obvious, it makes the algorithm to converge linearly, while algorithms without are saturating at a high level. This stress on the importance of using the memory in non-i.i.d. settings.



(a) LR: $\sigma_*^2 = 0$

(b) X-axis in # bits.

Figure S11: **Polyak-Ruppert averaging, synthetic dataset.** Logistic Regression on non-i.i.d. data using a batch gradient descent (to get $\sigma_* = 0$) and a Polyak-Ruppert averaging. The convergence is linear as predicted by Theorem 2 because $\sigma_* = 0$. Best seen in colors.

C.2 Real dataset: *superconduct* and *quantum*

In this section, we present details about experiments done on real-life datasets: *superconduct* (from Caruana et al.) which use a least-square regression, and *quantum* (from Hamidieh) with a logistic regression.

There is $N = 20$ devices (except for experiments using devices random sampling ; they use $N = 1000$ devices) for *superconduct* and *quantum* dataset. For *superconduct*, there is between 250 and 3900 points by worker, with a median at 750 ; and for *quantum*, there is between 900 and 10500 points, with a median at 2300. On each figure, we indicate which step size γ has been used.

Figures S12b and S13b corresponds to Figure 6 and are compared to the case when $b = 1$. Figure 7 is reproduced on Figure S14 to compare with the case when there is a partial participation (and $p = 0.5$, see Assumption 6).

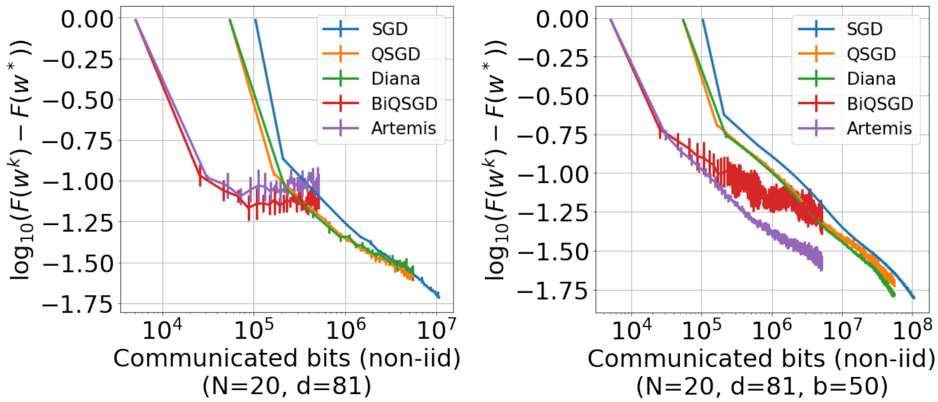
We observe that with a pure stochastic descent ($b = 1$), the memory mechanism does not really improve the convergence. But as soon as we increase the batch size (without however reaching a full batch gradient i.e $\sigma_* \neq 0$), the benefit of the memory becomes obvious ! In the case of the *quantum* dataset (see Figure S13), `Artemis` is not only better than `Bi-QSGD`, but also better than `QSGD`. **That is to say, we achieve to make an algorithm using a bidirectional compression, better than an algorithm handling a unidirectional compression.**

The comparison of the algorithm behaviors when $b = 1$ and $b > 1$, is an excellent illustration of Theorem 1 and of the impact of the noise term E (see Table 2). Indeed, while we increase the batch size, the noise σ_*^2/b decrease and thus has less impact on the convergence. Furthermore as the memory mechanism is making the terms in B^2 to be removed, the variance is much lower, which follows that the saturation level is also much lower.

C.2.1 Optimized step size

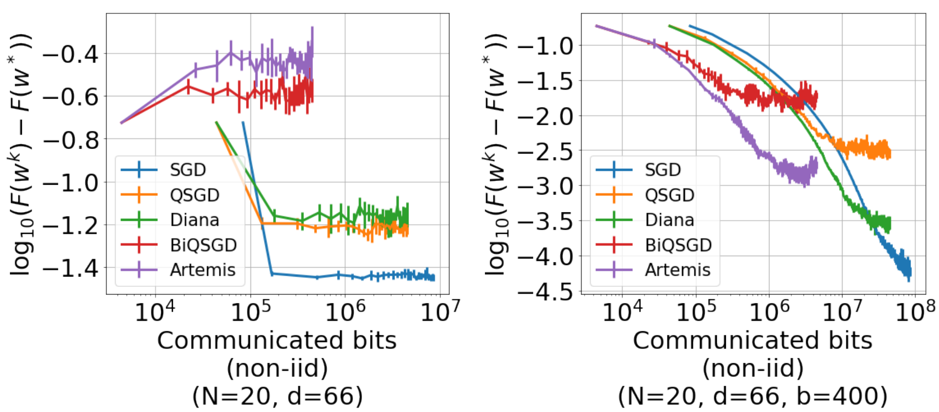
In this section, we want to address the issue of the optimal step size. On Figure S15 we plot the minimal loss after 250 iterations for each of the 5 algorithms, we can see that algorithms with memory clearly outperform those without. Then, on Figure S16 we present the loss of `Artemis` after 250 iterations 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 $[\![1, 250]\!]$ iterations. Finally, on Figure S17, we plot the loss obtain with the optimal step size γ_{opt} of each algorithms (found with Figure S15) w.r.t the number of communicated bits.

This last figures 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`.



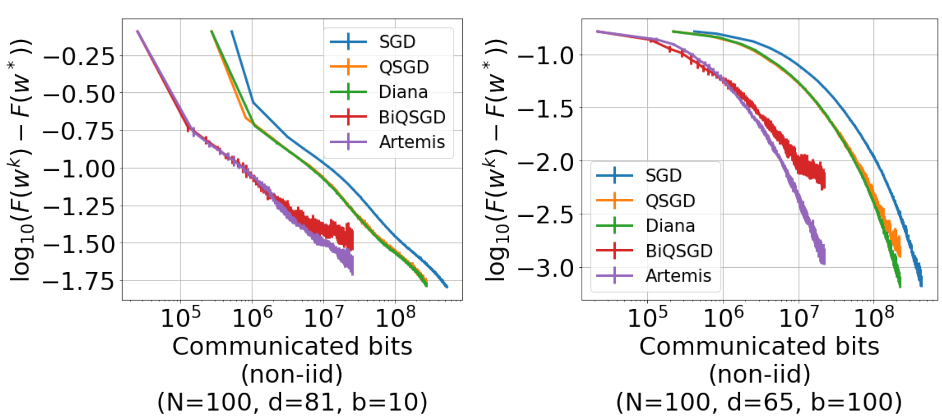
(a) Batch size $b = 1$ (b) Batch size $b = 50$

Figure S12: ***Superconduct***. LSR, $\sigma_* \neq 0$, non-i.i.d., X-axis in # bits, $\gamma = \frac{1}{8L}$. While we are increasing the batch size b, the noise is reduced, making the benefit of memory more obvious, this is a good illustration of Theorem 1 (see Table 2). Best seen in colors.



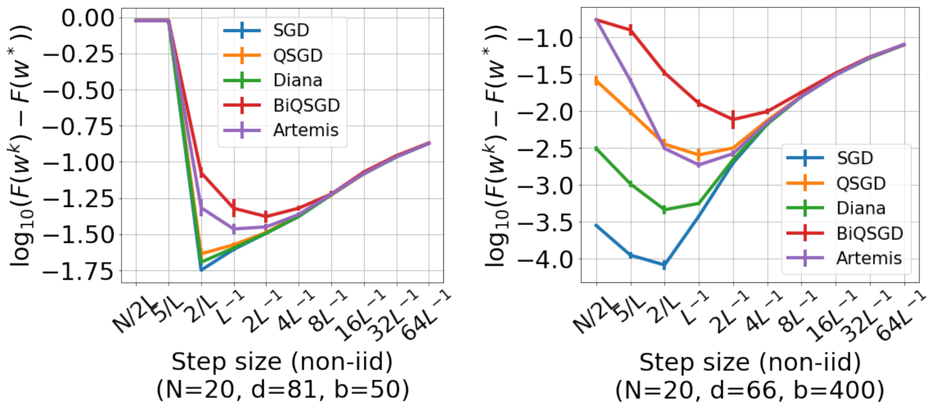
(a) Batch size $b = 1$ (b) Batch size $b = 400$

Figure S13: ***Quantum***. LR, non-i.i.d., $\sigma_* \neq 0$, X-axis in # bits, $\gamma = \frac{1}{8L}$. For $b = 1$, algorithms with bidirectional compression do not succeed to converge, but when the batch size is increased to 400, the advantage of memory becomes clear. This illustrate Theorem 1 (see Table 2). Best seen in colors.



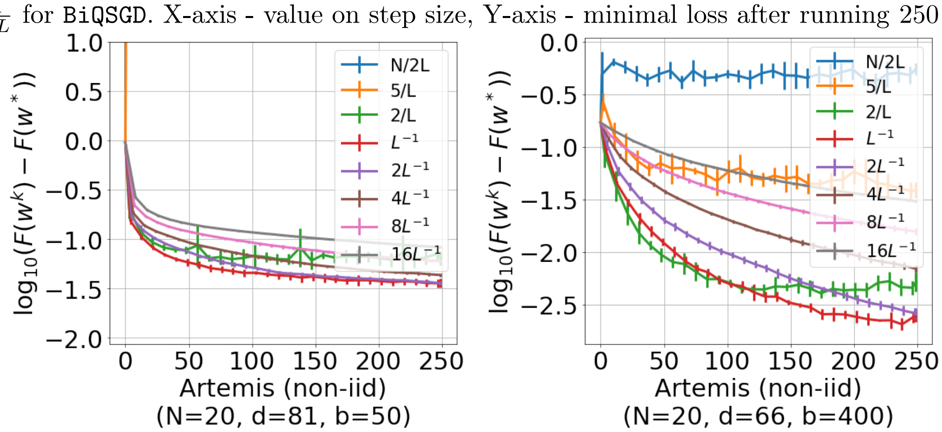
(a) Superconduct (b) Quantum

Figure S14: **Devices partial participation**: $\sigma_* \neq 0$, $\gamma = \frac{1}{L}$, $N = 100$ devices and $p = 0.5$ (1000 iter.). X-axis in # bits. This illustrates that `Artemis` correctly handles the case where only a subsample of devices are active. Best seen in colors.



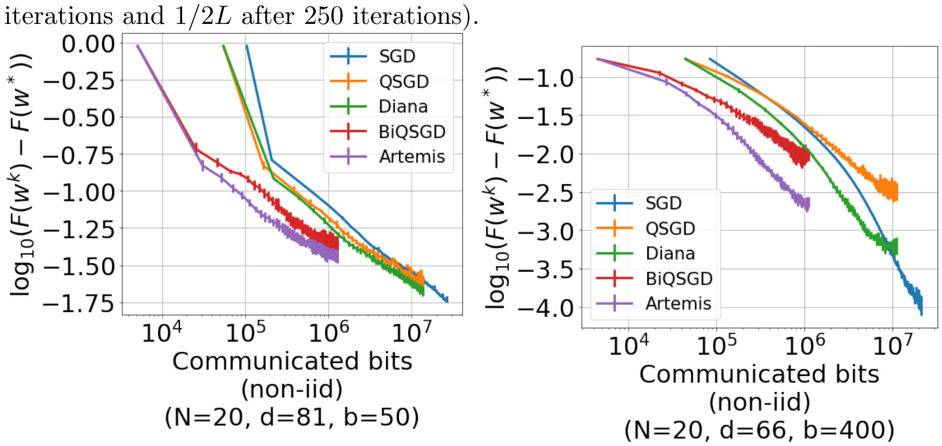
(a) Superconduct (b) Quantum

Figure S15: **Searching for the optimal step size γ_{opt} for each algorithm.** It is interesting to note that the memory allows to attain the lowest loss with a bigger step size. So, the optimal step size is $\gamma_{opt} = \frac{1}{L}$ for `Artemis` , but is $\gamma_{opt} = \frac{1}{2L}$ for `BiQSGD`. X-axis - value on step size, Y-axis - minimal loss after running 250 iterations.



(a) Superconduct (b) Quantum

Figure S16: **Loss w.r.t. step size γ.** Plotting the loss of `Artemis` after 250 iterations for different step size. As stressed by Figure S15, after 250 iteration, the best accuracy is indeed obtained for the both dataset with $\gamma_{opt} = \frac{1}{L}$ for `Artemis`, and the optimal step size decreases with the number of iterations (e.g., for *quantum*, it is $2/L$ before 100 iterations and $1/2L$ after 250 iterations).



(a) Superconduct

(b) Quantum

Figure S17: **Optimal step size.** X-axis in # bits. Plotting the loss of each algorithm obtained with the optimal gamma that attain the lowest error after 250 iteration (for instance $\gamma = \frac{1}{L}$ for `Artemis`, but $\gamma = \frac{2}{L}$ for `SGD`). For both *superconduct* and *quantum* dataset, taking the optimal step size of each algorithm, `Artemis` is superior to other variants w.r.t. accuracy and number of bits.

C.3 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. [2020]. 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.

D Technical Results

In this section, we introduce a few technical lemmas that will be used in the proofs. In Appendix D.1, we give four simple lemmas, while in Appendix D.2 we present a lemma which will be invoked in Appendix E to demonstrate Theorems S4 to S6.

Notation. Let $k^* \in \mathbb{N}$ and $(a_{k+1}^i)_{i=0}^N \in (\mathbb{R}^d)^N$ random variables independent of each other and $\mathcal{J}_{k+1}$-measurable, for a σ-field $\mathcal{J}_{k+1}$ s.t. $(B_k^i)_{i=1}^N$ are independent of $\mathcal{J}_{k+1}$.. Then in all the following demonstration, we note $a_{k+1} = \frac{1}{N}\sum_{i=1}^N a_{k+1}^i$ and $a_{k+1,S_k} = \frac{1}{pN}\sum_{i\in S_k} a_{k+1}^i$.

The vector a_{k+1} (resp. the σ-field $\mathcal{J}_{k+1}$) may represent various objects, for instance : g_{k+1}, $\widehat{g}_{k+1}$, G_{k+1} (resp. $\mathcal{F}_{k+1}$, $\mathcal{G}_{k+1}$, $\mathcal{H}_{k+1}$).

Remarks on the assumptions. We can add the following remarks on the assumptions.

- Assumption 6 can be extended to probabilities depending on the worker $(p_i)_{i\in[\![1,N]\!]}$.
- Assumption 5 requires in fact to access a *sequence of i.i.d. compression operators* $\mathcal{C}_{\mathrm{up/dwn},k}$ for $k \in \mathbb{N}$ – but for simplicity, we generally omit the k index.
- Assumption 4 in fact only requires that for any $i \in [\![1, N]\!]$, $\mathbb{E}\left[\|g_{k+1,*}^i - \nabla F_i(w_*)\|^2 \,\middle|\, \mathcal{H}_k\right] \leq \frac{(\sigma_*)_i^2}{b}$, and the results then hold for $\sigma_* = \frac{1}{N}\sum_{i=1}^N (\sigma_*)_i^2$. In other words, the bounds does not need to be uniform over workers, only the average truly matters.

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:*

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

This inequality is a consequence of strong convexity and can be found in [Nesterov, 2004, equation 2.1.22].

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 6 are verified.

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 the compressed term sent by each node to the central server can be bounded as following:*

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

Proof. Let $k \in \mathbb{N}$ and $i \in [\![1, N]\!]$, we have by definition:

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

Applying Lemma S1 gives the expected result. □

The aim of Lemma S6 is to express the variance of $\widehat{\Delta}_k = \sum_{i \in S_k} \mathcal{C}_{\mathrm{up}}(\Delta_k^i)$ making appear each term Δ_k^i, these terms will be later bounded with Lemma S5.

Lemma S6 (Variance of $\widehat{\Delta}_k$)**.** *The variance of the sum of the squared norm terms sent by each active node verifies:*

$$\forall k \in \mathbb{N}, \quad \mathbb{E}\left[\left\|\widehat{\Delta}_{k,S_k} - \Delta_{k,S_k}\right\|^2 \,\middle|\, \sigma(\mathcal{F}_{k+1} \cup \mathcal{B}_k)\right] \leq \frac{\omega_{\mathcal{C}}^{\mathrm{up}}}{p^2 N^2} \sum_{i \in S_k} \left\|\Delta_k^i\right\|^2\,.$$

Proof. Let $k \in \mathbb{N}$ and $i \in [\![1, N]\!]$, we have by definition:

$$\begin{aligned}\left\|\widehat{\Delta}_{k,S_k} - \Delta_{k,S_k}\right\|^2 &= \left\|\frac{1}{pN}\sum_{i \in S_k} \widehat{\Delta}_k^i - \Delta_k^i\right\|^2 \text{ and developing the squared norm,}\\ &= \frac{1}{p^2N^2}\sum_{i \in S_k}\left\|\widehat{\Delta}_k^i - \Delta_k^i\right\|^2 + \frac{1}{pN}\sum_{i,j \in S_k / i \neq j} \underbrace{\left\langle \widehat{\Delta}_k^i - \Delta_k^i \,\middle|\, \widehat{\Delta}_k^j - \Delta_k^j\right\rangle}_{=0 \text{ in expectation by independence of } (\widehat{\Delta}_k^i)_{i=1}^N}\,,\end{aligned}$$

and recall that $\mathbb{E}\left[\left\|\widehat{\Delta}_k^i - \Delta_k^i\right\|^2 \,\middle|\, \sigma(\mathcal{F}_{k+1} \cup \mathcal{B}_k)\right] \leq \omega_{\mathcal{C}}^{\mathrm{up}}\left\|\Delta_k^i\right\|^2$ (Proposition S4) gives the result. □

Below, we show up a recursion over the memory term h_k^i involving the stochastic gradients. This recursion will be used in Lemma S14. The existence of recursion has been first shed into light by Mishchenko et al. [2019].

Lemma S7 (Expectation of memory term)**.** *The memory term h_{k+1}^i can be expressed using a recursion involving the stochastic gradient g_{k+1}^i:*

$$\forall k \in \mathbb{N}\,,\, \forall i \in [\![1, N]\!]\,, \quad \mathbb{E}\left[h_{k+1}^i \,\middle|\, \mathcal{F}_{k+1}\right] = (1-\alpha)h_k^i + \alpha g_{k+1}^i\,.$$

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

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

and considering that $\mathbb{E}\left[\widehat{g}_{k+1}^i \,\middle|\, \mathcal{F}_{k+1}\right] = g_{k+1}^i$ (Proposition S3), the proof is completed.

□

In Lemma S8, we rewrite $\|g_{k+1}\|^2$ and $\|g_{k+1} - h_*^i\|^2$ to make appears:

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

Lemma S8 is required to correctly apply cocoercivity in Lemma S15.

Lemma S8 (Before using co-coercivity)**.** *Let $k \in [\![0, K]\!]$ and $i \in [\![1, N]\!]$. The noise in the stochastic gradients as defined in Assumptions 3 and 4 can be controlled as following:*

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_{k+1}^i\right\|^2 \;\middle|\; \mathcal{H}_k\right] \leq \frac{2}{N}\sum_{i=1}^{N}\left(\mathbb{E}\left[\left\|g_{k+1}^i - g_{k+1,*}^i\right\|^2 \;\middle|\; \mathcal{H}_k\right] + \left(\frac{\sigma_*^2}{b} + B^2\right)\right), \tag{S4}$$

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_{k+1}^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_k\right] \leq \frac{2}{N}\sum_{i=1}^{N}\left(\mathbb{E}\left[\left\|g_{k+1}^i - g_{k+1,*}^i\right\|^2 \;\middle|\; \mathcal{H}_k\right] + \frac{\sigma_*^2}{b}\right). \tag{S5}$$

Proof. Let $k \in \mathbb{N}$. For eq. (S4):

$$\begin{aligned}\left\|g_{k+1}^i\right\|^2 &= \left\|g_{k+1}^i - g_{k+1,*}^i + g_{k+1,*}^i\right\|^2 \\ &\leq 2\left(\left\|g_{k+1}^i - g_{k+1,*}^i\right\|^2 + \left\|g_{k+1,*}^i\right\|^2\right) \text{ using inequality of Lemma S1.}\end{aligned}$$

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

For eq. (S5), we use Lemma S1 and we write:

$$\begin{aligned}\left\|g_{k+1}^i - h_*^i\right\|^2 &= \left\|\left(g_{k+1}^i - g_{k+1,*}^i\right) + \left(g_{k+1,*}^i - \nabla F_i(w_*)\right)\right\|^2 \\ &\leq 2\left(\left\|g_{k+1}^i - g_{k+1,*}^i\right\|^2 + \left\|g_{k+1,*}^i - \nabla F_i(w_*)\right\|^2\right).\end{aligned}$$

Taking expectation, we have:

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

□

Lemma S9 is used in the memory case, it helps to pass for k in $\mathbb{N}$ from $\widehat{g}_{k+1,S_k}$ to $(\widehat{g}_{k+1}^i)_{i=1}^N$, this lemma will allow to invoke Lemma S10.

Lemma S9. *In the case without memory, we have the following bound on the squared norm of the compressed gradient (randomly sampled), for all k in $\mathbb{N}$:*

$$\mathbb{E}\left[\|\widehat{g}_{k+1,S_k}\|^2 \;\middle|\; \mathcal{G}_{k+1}\right] = \frac{(1-p)}{pN^2}\sum_{i=0}^{N}\left\|\widehat{g}_{k+1}^i\right\|^2 + \|\widehat{g}_{k+1}\|^2 .$$

Proof. The proof is quite straightforward using the expectation and the variance of $\widehat{g}_{k+1,S_k}$ computed in Propositions S7 and S8. We just need to decompose as following to easily obtain the result:

$$\mathbb{E}\left[\|\widehat{g}_{k+1,S_k}\|^2 \;\middle|\; \mathcal{G}_{k+1}\right] = \mathbb{V}\left[\widehat{g}_{k+1,S_k} \mid \mathcal{G}_{k+1}\right] + \left\|\mathbb{E}\left[\widehat{g}_{k+1,S_k} \mid \mathcal{G}_{k+1}\right]\right\|^2$$

□

Lemma S10 is used in the case without memory. It is used to remove the uplink compression noise straight after Lemma S9 has been applied.

Lemma S10 (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 (randomly sampled), for all k in $\mathbb{N}$:*

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

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

$$\begin{aligned}
\|\widehat{g}_{k+1}\|^2 &= \|\widehat{g}_{k+1} - g_{k+1} + g_{k+1}\|^2 \\
&= \|\widehat{g}_{k+1} - g_{k+1}\|^2 + 2 \left\langle \widehat{g}_{k+1} - g_{k+1} \mid g_{k+1} \right\rangle + \|g_{k+1}\|^2 .
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[\|\widehat{g}_{k+1}\|^2 \;\middle|\; \mathcal{F}_{k+1}\right] \;\middle|\; \mathcal{H}_k\right] = \mathbb{E}&\left[\mathbb{E}\left[\|\widehat{g}_{k+1} - g_{k+1}\|^2 \;\middle|\; \mathcal{F}_{k+1}\right] \;\middle|\; \mathcal{H}_k\right] \\
&+ 2 \times \mathbb{E}\left[\mathbb{E}\left[\left\langle \widehat{g}_{k+1} - g_{k+1} \mid g_{k+1} \right\rangle \mid \mathcal{F}_{k+1}\right] \mid \mathcal{H}_k\right] \\
&+ \mathbb{E}\left[\|g_{k+1}\|^2 \;\middle|\; \mathcal{H}_k\right] .
\end{aligned} \tag{S6}$$

We need to find a bound for each of the terms of above eq. (S6). The last term is handled in Lemma S12, narrowing it down to the case $p = 1$.

It follows that we just need to bound $\|\widehat{g}_{k+1} - g_{k+1}\|^2$:

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

Combining with Proposition S4, we hold that:

$$\mathbb{E}\left[\|\widehat{g}_{k+1} - g_{k+1}\|^2 \;\middle|\; \mathcal{F}_{k+1}\right] \leq \frac{\omega_{\mathcal{C}}^{\mathrm{up}}}{N^2} \sum_{i=1}^{N} \left\|g_{k+1}^i\right\|^2 .$$

Now, we proved that:

$$\begin{cases}
\mathbb{E}\left[\|\widehat{g}_{k+1} - g_{k+1}\|^2 \;\middle|\; \mathcal{F}_{k+1}\right] & \leq \frac{\omega_{\mathcal{C}}^{\mathrm{up}}}{N^2} \sum_{i=1}^{N} \mathbb{E}\left[\left\|g_{k+1}^i\right\|^2 \;\middle|\; \mathcal{F}_{k+1}\right] \\
\mathbb{E}\left[\left\langle \widehat{g}_{k+1} - g_{k+1} \mid g_{k+1} \right\rangle \mid \mathcal{F}_{k+1}\right] & = 0 \qquad \text{(Proposition S3)} \\
\mathbb{E}\left[\|g_{k+1}\|^2 \;\middle|\; \mathcal{H}_k\right] & \leq \frac{1}{N^2} \sum_{i=0}^{N} \mathbb{E}\left[\left\|g_{k+1}^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_k\right] \\
& \quad + L \left\langle \nabla F(w_k) \mid w_k - w_* \right\rangle \quad (\text{ Lemma S12, with } p = 1) .
\end{cases}$$

Thus, we obtain from eq. (S6):

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

□

Lemma S11. *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 $[\![1,N]\!]$:* $\mathbb{E}\left[\left\|\widehat{g}_{k+1}^i\right\|^2 \;\middle|\; \mathcal{F}_{k+1}\right] \leq (\omega_{\mathcal{C}}^{\mathrm{up}}+1)\left\|g_{k+1}^i\right\|^2$

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

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{g}_{k+1}^i\right\|^2 \;\middle|\; \mathcal{F}_{k+1}\right] &= \mathbb{E}\left[\left\|\widehat{g}_{k+1}^i - g_{k+1}^i + g_{k+1}^i\right\|^2 \;\middle|\; \mathcal{F}_{k+1}\right] \\
&= \mathbb{E}\left[\left\|\widehat{g}_{k+1}^i - g_{k+1}^i\right\|^2 \;\middle|\; \mathcal{F}_{k+1}\right] + 2\underbrace{\mathbb{E}\left[\left\langle \widehat{g}_{k+1}^i - g_{k+1}^i \mid g_{k+1}^i\right\rangle \;\middle|\; \mathcal{F}_{k+1}\right]}_{=0} + \mathbb{E}\left[\left\|g_{k+1}^i\right\|^2 \;\middle|\; \mathcal{F}_{k+1}\right]
\end{aligned}$$

To obtain the result, we need to recall that $\left\|g_{k+1}\right\|^2$ is $\mathcal{F}k+1$-measurable, and then to use Proposition S6. □

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

Lemma S12. *Regardless if we use memory, we have the following bound on the squared norm of the gradient, for all k in $\mathbb{N}$:*

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

Proof. Let $k \in \mathbb{N}$,

$$\begin{aligned}
\left\|g_{k+1,S_k}\right\|^2 &= \left\|\frac{1}{pN}\sum_{i\in S_k} g_{k+1}^i\right\|^2 \\
&= \left\|\frac{1}{pN}\sum_{i\in S_k}\left(g_{k+1}^i - \nabla F_i(w_k)\right) + \frac{1}{pN}\sum_{i\in S_k}\nabla F_i(w_k)\right\|^2 .
\end{aligned}$$

Now taking conditional expectation w.r.t $\sigma(\mathcal{I}_k \cup \mathcal{B}_k)$ (including $\mathcal{B}_k$ in the σ-field allows to not consider the randomness associated to the device sampling):

$$\mathbb{E}\left[\left\|g_{k+1,S_k}\right\|^2 \;\middle|\; \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] = \mathbb{E}\left[\left\|\frac{1}{pN}\sum_{i\in S_k} g_{k+1}^i - \nabla F_i(w_k) + \frac{1}{pN}\sum_{i\in S_k}\nabla F_i(w_k)\right\|^2 \;\middle|\; \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] .$$

Expanding this squared norm:

$$\begin{aligned}
\mathbb{E}\left[\left\|g_{k+1,S_k}\right\|^2 \;\middle|\; \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] &= \mathbb{E}\left[\left\|\frac{1}{pN}\sum_{i\in S_k} g_{k+1}^i - \nabla F_i(w_k)\right\|^2 \;\middle|\; \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] \\
&\quad + 2\mathbb{E}\left[\left\langle \frac{1}{pN}\sum_{i\in S_k} g_{k+1}^i - \nabla F_i(w_k) \;\middle|\; \frac{1}{pN}\sum_{i\in S_k}\nabla F_j(w_k)\right\rangle \;\middle|\; \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] \\
&\quad + \mathbb{E}\left[\left\|\frac{1}{pN}\sum_{i\in S_k}\nabla F_i(w_k)\right\|^2 \;\middle|\; \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] .
\end{aligned}$$

Moreover, $\forall i, j \in [\![1, N]\!]^2, \mathbb{E}\left[\left\langle g_{k+1}^i - \nabla F_i(w_k) \;\middle|\; \nabla F_j(w_k)\right\rangle \;\middle|\; \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] = 0$ and $\nabla F(w_k)$ is $\sigma(\mathcal{I}_k \cup \mathcal{B}_k)$-measurable:

$$\mathbb{E}\left[\|g_{k+1,S_k}\|^2 \;\middle|\; \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] \leq \mathbb{E}\left[\left\|\frac{1}{pN}\sum_{i\in S_k} g_{k+1}^i - \nabla F_i(w_k)\right\|^2 \;\middle|\; \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] + \|\nabla F_{s_k}(w_k)\|^2 \,. \tag{S7}$$

To compute $\|\nabla F_{S_k}(w_k)\|^2$, we apply cocoercivity (see Assumption 2) and next we take expectation w.r.t σ-algebra $\mathcal{I}_k$:

$$\mathbb{E}\left[\|\nabla F_{S_k}(w_k)\|^2 \;\middle|\; \mathcal{I}_k\right] = L \left\langle \mathbb{E}\left[\nabla F_{S_k}(w_k) \mid \mathcal{I}_k\right] \;\middle|\; w_k - w_*\right\rangle = L \left\langle \nabla F(w_k) \mid w_k - w_*\right\rangle$$

Now, for sake of clarity we note $\Pi = \left\|\frac{1}{pN}\sum_{i\in S_k} g_{k+1}^i - \nabla F_i(w_k)\right\|^2$, then:

$$\begin{aligned}
\mathbb{E}\left[\Pi \mid \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] &= \frac{1}{p^2N^2}\sum_{i\in S_k}\mathbb{E}\left[\left\|g_{k+1}^i - \nabla F_i(w_k)\right\|^2 \;\middle|\; \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] \\
&\quad + \frac{1}{p^2N^2}\sum_{i,j\in S_k/i\neq j}\underbrace{\mathbb{E}\left[\left\langle g_{k+1}^i - \nabla F_i(w_k) \;\middle|\; g_{k+1}^j - \nabla F_j(w_k)\right\rangle \;\middle|\; \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right]}_{=0 \text{ by independence of } (g_{k+1}^i)_{i=0}^N} \\
&= \frac{1}{p^2N^2}\sum_{i\in S_k}\mathbb{E}\left[\left\|(g_{k+1}^i - \nabla F_i(w_*)) + (\nabla F_i(w_*) - \nabla F_i(w_k))\right\|^2 \;\middle|\; \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right].
\end{aligned}$$

Developing the squared norm a second time:

$$\begin{aligned}
\mathbb{E}\left[\Pi \mid \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] &= \frac{1}{p^2N^2}\sum_{i\in S_k}\mathbb{E}\left[\left\|g_{k+1}^i - \nabla F_i(w_*)\right\|^2 \;\middle|\; \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] \\
&\quad + \frac{2}{p^2N^2}\sum_{i\in S_k}\mathbb{E}\left[\left\langle g_{k+1}^i - \nabla F_i(w_*) \;\middle|\; \nabla F_i(w_*) - \nabla F_i(w_k)\right\rangle \;\middle|\; \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] \\
&\quad + \frac{1}{p^2N^2}\sum_{i\in S_k}\|\nabla F_i(w_k) - \nabla F_i(w_*),\|^2
\end{aligned}$$

Then,

$$\begin{aligned}
\mathbb{E}\left[\Pi \mid \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] &= \frac{1}{p^2N^2}\sum_{i\in S_k}\mathbb{E}\left[\left\|g_{k+1}^i - \nabla F_i(w_*)\right\|^2 \;\middle|\; \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] \\
&\quad - \frac{2}{p^2N^2}\sum_{i\in S_k}\left\langle \nabla F_i(w_k) - \nabla F_i(w_*) \mid \nabla F_i(w_k) - \nabla F_i(w_*)\right\rangle \\
&\quad + \frac{1}{p^2N^2}\sum_{i\in S_k}\|\nabla F_i(w_k) - \nabla F_i(w_*)\|^2 \\
&= \frac{1}{p^2N^2}\sum_{i\in S_k}\mathbb{E}\left[\left\|g_{k+1}^i - \nabla F_i(w_*)\right\|^2 \;\middle|\; \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] - \|\nabla F_i(w_k) - \nabla F_i(w_*)\|^2
\end{aligned}$$

applying cocoercivity (Assumption 2):

$$\mathbb{E}\left[\Pi \mid \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] \leq \frac{1}{p^2N^2}\sum_{i\in S_k}\mathbb{E}\left[\left\|g_{k+1}^i - \nabla F_i(w_*)\right\|^2 \;\middle|\; \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] - L\left\langle \nabla F_i(w_k) - \nabla F_i(w_*) \mid w_k - w_*\right\rangle. \tag{S8}$$

Now we consider the randomness associated to device sampling. Remember that because $\mathcal{I}_k \subset \sigma(\mathcal{I}_k \cup \mathcal{B}_k)$, we have $\mathbb{E}\left[\Pi \mid \mathcal{I}_k\right] = \mathbb{E}\left[\mathbb{E}\left[\Pi \mid \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] \mid \mathcal{I}_k\right]$. Thus, we consider now Π w.r.t. the σ-field $\mathcal{I}_k$

Handling first term of eq. (S8):

$$\frac{1}{p^2N^2} \sum_{i \in S_k} \mathbb{E}\left[\left\| g_{k+1}^i - \nabla F_i(w_*) \right\|^2 \;\middle|\; \mathcal{I}_k \right] = \frac{1}{p^2N^2} \sum_{i=1}^N B_k^i \mathbb{E}\left[\left\| g_{k+1}^i - \nabla F_i(w_*) \right\|^2 \;\middle|\; \mathcal{I}_k \right]$$
$$= \frac{1}{pN^2} \sum_{i=1}^N \mathbb{E}\left[\left\| g_{k+1}^i - \nabla F_i(w_*) \right\|^2 \;\middle|\; \mathcal{I}_k \right] .$$

Handling second term of eq. (S8):

$$L \left\langle \frac{1}{p^2N^2} \sum_{i \in S_k} \nabla F_i(w_k) - \nabla F_i(w_*) \;\middle|\; w_k - w_* \right\rangle = L \left\langle \frac{1}{p^2N^2} \sum_{i=0}^N \left(\nabla F_i(w_k) - \nabla F_i(w_*) \right) B_k^i \;\middle|\; w_k - w_* \right\rangle .$$

Taking expectation w.r.t σ-algebra $\mathcal{I}_k$:

$$\mathbb{E}\left[L \left\langle \frac{1}{p^2N^2} \sum_{i \in S_k} \nabla F_i(w_k) - \nabla F_i(w_*) \;\middle|\; w_k - w_* \right\rangle \;\middle|\; \mathcal{I}_k \right]$$
$$= L \left\langle \frac{1}{pN^2} \sum_{i=0}^N \nabla F_i(w_k) - \nabla F_i(w_*) \;\middle|\; w_k - w_* \right\rangle$$
$$= \frac{L}{pN} \left\langle \nabla F(w_k) - \nabla F_i(w_*) \mid w_k - w_* \right\rangle .$$

Injecting this in eq. (S8):

$$\mathbb{E}\left[\Pi \mid \mathcal{I}_k \right] \leq \frac{}{pN^2} \sum_{i=1}^N \mathbb{E}\left[\left\| g_{k+1}^i - \nabla F_i(w_*) \right\|^2 \;\middle|\; \mathcal{I}_k \right] - \frac{L}{pN} \left\langle \nabla F(w_k) \mid w_k - w_* \right\rangle .$$

Recall that we note $h_*^i = \nabla F_i(w_*)$, returning to eq. (S7) and invoking again cocoercivity:

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

which we simplify by considering that:

$$\mathbb{E}\left[\left\| g_{k+1} \right\|^2 \;\middle|\; \mathcal{I}_k \right] \leq \frac{1}{pN^2} \sum_{i=1}^N \mathbb{E}\left[\left\| g_{k+1}^i - h_*^i \right\|^2 \;\middle|\; \mathcal{I}_k \right] + L \left\langle \nabla F(w_k) \mid w_k - w_* \right\rangle .$$

□

In order to derive an upper bound on the squared norm of $\|w_{k+1} - w_*\|^2$, for k in $\mathbb{N}$, we need to control $\|\widehat{g}_{k+1,S_k}\|^2$. This term is decomposed as a sum of three terms depending on:

1. the recursion over the memory term (h_k^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 S13. *In the case with memory, we have the following upper bound on the squared norm of the compressed gradient, for all k in $\mathbb{N}$:*

$$\mathbb{E}\left[\left\|\widehat{g}_{k+1,S_k}\right\|^2 \;\middle|\; \mathcal{I}_k\right] \leq \frac{2(2\omega_{\mathcal{C}}^{\mathrm{up}}+1)}{pN^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_{k+1}^i - g_{k+1,*}^i\right\|^2 \;\middle|\; \mathcal{I}_k\right] + \frac{2\omega_{\mathcal{C}}^{\mathrm{up}}}{pN^2}\sum_{i=1}^N \mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{I}_k\right] + L\left\langle \nabla F(w_k) \mid w_k - w_*\right\rangle + \frac{2(2\omega_{\mathcal{C}}^{\mathrm{up}}+1)}{pN} \times \frac{\sigma_*^2}{b}\,.$$

Proof. Let k in $\mathbb{N}$. First, because $\sigma(\mathcal{I}_k \cup \mathcal{B}_k) \subset \sigma(\mathcal{F}_{k+1} \cup \mathcal{B}_k)$, we have:

$$\mathbb{E}\left[\widehat{g}_{k+1,S_k} \mid \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right] = \mathbb{E}\left[\mathbb{E}\left[\widehat{g}_{k+1,S_k} \mid \sigma(\mathcal{F}_{k+1} \cup \mathcal{B}_k)\right] \mid \sigma(\mathcal{I}_k \cup \mathcal{B}_k)\right]\,.$$

Then,

$$\begin{aligned}\left\|\widehat{g}_{k+1,S_k}\right\|^2 &= \left\|\widehat{\Delta}_{k,S_k} + h_k\right\|^2 = \left\|(\widehat{\Delta}_{k,S_k} - \Delta_{k,S_k}) + (h_k + \Delta_{k,S_k})\right\|^2 \\ &= \left\|\widehat{\Delta}_{k,S_k} - \Delta_{k,S_k}\right\|^2 + \left\langle \widehat{\Delta}_{k,S_k} - \Delta_{k,S_k} \;\middle|\; h_k + \Delta_{k,S_k}\right\rangle + \left\|h_k + \Delta_{k,S_k}\right\|^2\,.\end{aligned}$$

Recall that $h_k + \Delta_{k,S_k} = g_{k+1,S_k}$, and taking expectation gives:

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{g}_{k+1,S_k}\right\|^2 \;\middle|\; \sigma(\mathcal{F}_{k+1} \cup \mathcal{B}_k)\right] &= \mathbb{E}\left[\left\|\widehat{\Delta}_{k,S_k} - \Delta_{k,S_k}\right\|^2 \;\middle|\; \sigma(\mathcal{F}_{k+1} \cup \mathcal{B}_k)\right] \\ &\quad + \mathbb{E}\left[\left\|g_{k+1,S_k}\right\|^2 \;\middle|\; \sigma(\mathcal{F}_{k+1} \cup \mathcal{B}_k)\right]\,, \text{ using Lemma S6:} \\ &\leq \frac{\omega_{\mathcal{C}}^{\mathrm{up}}}{p^2N^2}\sum_{i\in S_k}\left\|\Delta_k^i\right\|^2 + \mathbb{E}\left[\left\|g_{k+1,S_k}\right\|^2 \;\middle|\; \sigma(\mathcal{F}_{k+1} \cup \mathcal{B}_k)\right]\,.\end{aligned}$$

Now, we consider the device sampling randomness and take expectation w.r.t $\mathcal{I}_k \subset \sigma(\mathcal{F}_{k+1} \cup \mathcal{B}_k)$ (the inclusion is true because $\mathcal{F}_{k+1}$ contains $\mathcal{B}_{k-1}$), we have that:

$$\mathbb{E}\left[\sum_{i\in S_k}\left\|\Delta_k^i\right\|^2 \;\middle|\; \mathcal{I}_k\right] = \mathbb{E}\left[\sum_{i=1}^N B_k^i\left\|\Delta_k^i\right\|^2 \;\middle|\; \mathcal{I}_k\right] = \sum_{i=1}^N p\left\|\Delta_k^i\right\|^2\,.$$

Using Lemma S5 to get rid of $(\left\|\Delta_k^i\right\|^2)_{i=1}^N$:

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{g}_{k+1,S_k}\right\|^2 \;\middle|\; \mathcal{I}_k\right] &\leq \frac{2\omega_{\mathcal{C}}^{\mathrm{up}}}{pN^2}\sum_{i\in S_k}\left(\mathbb{E}\left[\left\|g_{k+1}^i - h_*^i\right\|^2 + \left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{I}_k\right]\right) \\ &\quad + \mathbb{E}\left[\left\|g_{k+1,S_k}\right\|^2 \;\middle|\; \mathcal{I}_k\right]\,.\end{aligned} \tag{S9}$$

Back to eq. (S9) and with the use of Lemma S12:

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

Now with eq. (S5) of Lemma S8

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

Which leads to:

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

which conclude the proof. □

To show that the Lyapunov function is a contraction, we need to find a bound for each terms. Bounding $\|w_{k+1} - w_*\|^2$, for k in $\mathbb{N}$, flows from update schema (see eq. (3)) decomposition. However the memory term $\left\|h_{k+1}^i - h_*^i\right\|^2$ involved in the Lyapunov function doesn't show up naturally.

The aim of Lemma S14 is precisely to provide a recursive bound over the memory term to highlight the contraction. Like Lemma S7, the following lemma comes from Mishchenko et al. [2019].

Lemma S14 (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:*

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

Proof. The proof is done in two steps:

1. First, we povide a recursive bound on memory when the device is used for the update (i.e such that for k in $\mathbb{N}$, for i in $\{1, ...N\}$, $B_k^i = 1$)
2. Then, we generalize to the case with all i in $[\![1, N,]\!]$ regardless to if they are used at the round k.

First part. Let $k \in \mathbb{N}$ and let $i \in [\![1, N]\!]$ such that $B_k^i = 1$

$$\begin{aligned}\mathbb{E}\left[\left\|h_{k+1}^i - h_*^i\right\|^2 \;\middle|\; \mathcal{F}_{k+1}\right] = \left\|\mathbb{E}\left[h_{k+1}^i \mid \mathcal{F}_{k+1}\right] - h_*^i\right\|^2\\ + \mathbb{E}\left[\left\|h_{k+1}^i - \mathbb{E}\left[h_{k+1}^i \mid \mathcal{F}_{k+1}\right]\right\|^2 \;\middle|\; \mathcal{F}_{k+1}\right] \quad \text{using Lemma S3},\end{aligned}$$

and now with Lemma S7:

$$\begin{aligned}= \left\|(1-\alpha)h_k^i + \alpha g_{k+1}^i - h_*^i\right\|^2\\ + \mathbb{E}\left[\left\|h_{k+1}^i - \mathbb{E}\left[h_{k+1}^i \mid \mathcal{F}_{k+1}\right]\right\|^2 \;\middle|\; \mathcal{F}_{k+1}\right] .\end{aligned}$$

Now recall that $h_{k+1}^i = h_k^i + \alpha \widehat{\Delta}_k^i$ and $\mathbb{E}\left[\widehat{\Delta}_k^i \;\middle|\; \mathcal{F}_{k+1}\right] = \Delta_k^i$:

$$\begin{aligned}\mathbb{E}\left[\left\|h_{k+1}^i - h_*^i\right\|^2 \;\middle|\; \mathcal{F}_{k+1}\right] &= \left\|(1-\alpha)(h_k^i - h_*^i) + \alpha(g_{k+1}^i - h_*^i)\right\|^2 \\ &\quad + \alpha^2 \mathbb{E}\left[\left\|\widehat{\Delta}_k^i - \Delta_k^i\right\|^2 \;\middle|\; \mathcal{F}_{k+1}\right].\end{aligned}$$

Using Lemma S2 of Appendix D.1 and Proposition S4:

$$\begin{aligned}&\leq (1-\alpha)\left\|h_k^i - h_*^i\right\|^2 + \alpha\left\|g_{k+1}^i - h_*^i\right\|^2 \\ &\quad - \alpha(1-\alpha)\left\|h_k^i - g_{k+1}^i\right\|^2 + \alpha^2 \omega_{\mathcal{C}}^{\mathrm{up}}\left\|\Delta_k^i\right\|^2.\end{aligned}$$

Because $h_k^i - g_{k+1}^i = \Delta_k^i$:

$$\begin{aligned}\mathbb{E}\left[\left\|h_{k+1}^i - h_*^i\right\|^2 \;\middle|\; \mathcal{F}_{k+1}\right] &\leq (1-\alpha)\left\|h_k^i - h_*^i\right\|^2 + \alpha\left\|g_{k+1}^i - h_*^i\right\|^2 \\ &\quad + \alpha\left(\alpha(\omega_{\mathcal{C}}^{\mathrm{up}}+1) - 1\right)\left\|\Delta_k^i\right\|^2,\end{aligned}$$

and using Lemma S5:

$$\begin{aligned}&\leq (1-\alpha)\left\|h_k^i - h_*^i\right\|^2 + \alpha\left\|g_{k+1}^i - h_*^i\right\|^2 \\ &\quad + 2\alpha\left(\alpha(\omega_{\mathcal{C}}^{\mathrm{up}}+1) - 1\right)\left(\left\|h_k^i - h_*^i\right\|^2 + \left\|g_{k+1} - h_*^i\right\|^2\right) \\ &\leq \left(1 + 2\alpha^2\omega_{\mathcal{C}}^{\mathrm{up}} + 2\alpha^2 - 3\alpha\right)\left\|h_k^i - h_*^i\right\|^2 \\ &\quad + \alpha(2\alpha\omega_{\mathcal{C}}^{\mathrm{up}} + 2\alpha - 1)\left\|g_{k+1} - h_*^i\right\|^2.\end{aligned}$$

Finally using eq. (S5) of Lemma S8 and writing that:

$$\mathbb{E}\left[\left\|g_{k+1} - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_k\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|g_{k+1} - h_*^i\right\|^2 \;\middle|\; \mathcal{F}_{k+1}\right] \;\middle|\; \mathcal{H}_k\right] (\text{because } \mathcal{H}_k \subset \mathcal{F}_{k+1}) ,$$

we have:

$$\begin{aligned}\mathbb{E}\left[\left\|h_{k+1}^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_k\right] &\leq (1 + \underbrace{2\alpha^2\omega_{\mathcal{C}}^{\mathrm{up}} + 2\alpha^2 - 3\alpha}_{=T_1})\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_k\right] \\ &\quad + 2(\underbrace{2\alpha^2\omega_{\mathcal{C}}^{\mathrm{up}} + 2\alpha^2 - \alpha}_{T_2})\mathbb{E}\left[\left\|g_{k+1} - g_{k+1,*}\right\|^2 \;\middle|\; \mathcal{H}_k\right] \\ &\quad + \underbrace{2\,\frac{\sigma_*^2}{b}\left(2\alpha^2(\omega_{\mathcal{C}}^{\mathrm{up}}+1) - \alpha\right)}_{T_3},\end{aligned}$$

which conclude the first part of the proof. Now we take the general case with $\forall i \in [\![1, N]\!], B_k^i = 0$ or 1.

Second part. Let $k \in \mathbb{N}$ and let $i \in [\![1, N]\!]$.

To resume, if the device participate to the iteration k, we have

$$\mathbb{E}\left[\left\|h_{k+1}^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_k\right] \leq (1+T_1)\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_k\right] + 2T_2\mathbb{E}\left[\left\|g_{k+1} - g_{k+1,*}\right\|^2 \;\middle|\; \mathcal{H}_k\right] + 2T_3\,,$$

otherwise:

$$\mathbb{E}\left[\left\|h_{k+1}^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_k\right] = \mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_k\right].$$

In other words, for all i in $[\![1, N]\!]$:

$$\begin{aligned}
\mathbb{E}\left[\left\|h_{k+1}^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_k\right] &\leq (1+T_1)B_k^i \mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_k\right] + 2T_2 B_k^i \mathbb{E}\left[\left\|g_{k+1} - g_{k+1,*}\right\|^2 \;\middle|\; \mathcal{H}_k\right] + 2T_3 B_k^i \\
&\quad + (1-B_k^i)\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_k\right] \\
&\leq (1+T_1 B_k^i)\mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{H}_k\right] + 2T_2 B_k^i \mathbb{E}\left[\left\|g_{k+1} - g_{k+1,*}\right\|^2 \;\middle|\; \mathcal{H}_k\right] + 2T_3 B_k^i
\end{aligned}$$

Taking expectation w.r.t σ-algebra $\mathcal{I}_k$ gives the result.

□

After successfully invoking all previous lemmas, we will finally be able to use co-coercivity. Lemma S15 shows how Assumption 2 is used to do it. After this stage, proof will be continued by applying strong-convexity of F.

Lemma S15 (Applying co-coercivity)**.** *This lemma shows how to apply co-coercivity on stochastic gradients.*

$$\forall k \in \mathbb{N}\,, \quad \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\left\|g_{k+1}^i - g_{k+1,*}\right\|^2 \;\middle|\; \mathcal{H}_k\right] \leq L \left\langle \nabla F(w_k) \mid w_k - w_* \right\rangle .$$

Proof. Let $k \in \mathbb{N}$.

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\left\|g_{k+1}^i - g_{k+1,*}^i\right\|^2 \;\middle|\; \mathcal{H}_k\right] &\leq \frac{1}{N}\sum_{i=1}^N L \left\langle \mathbb{E}\left[g_{k+1}^i - g_{k+1,*}^i \mid \mathcal{H}_k\right] \mid w_k - w_* \right\rangle \text{using Assumption 2,} \\
&\leq L \left\langle \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[g_{k+1}^i - g_{k+1,*}^i \mid \mathcal{H}_k\right] \;\middle|\; w_k - w_* \right\rangle \\
&\leq L \left\langle \frac{1}{N}\sum_{i=1}^N \nabla F_i(w_k) - \nabla F_i(w_*) \;\middle|\; w_k - w_* \right\rangle .
\end{aligned}$$

□

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. [2019]. And in a second time, we give a complete demonstration of theorems stated in the main paper: Theorems 2 and 3.

For the sake of demonstration, we define a Lyapunov function V_k [as in Mishchenko et al., 2019; Liu et al., 2020], with k in $[\![1, K]\!]$ and p in $\mathbb{R}^*$:

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

The Lyapunov function is defined 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`, and also when using Polyak-Ruppert averaging. To show that it's a contraction, we need three stages:

1. we develop the update schema defined in eq. (3) to get a first bound on $\left\|w_k - w_*\right\|^2$

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.

Clarification on the two equivalent notations Ω_{k+1,S_k} and G_{k+1,S_k}. In this article, we use two notation - Ω_{k+1,S_k} and G_{k+1,S_k} - for the same object $\mathcal{C}_{\mathrm{dwn}}\left(\widehat{g}_{k+1,S_k}\right)$. The notation Ω stresses on the fact that this corresponds to the information compressed and communicated (like Δ) from central server to local workers. The notation G underlines that it also corresponds to a sum of gradients. If a memory was added for the downlink communication, the two objects would not be equal! This explain why we decided to use two notations: to highlight which aspect was predominant in each situation.

E.1 Proof of Theorem S4 - variant without memory of `Artemis`

Theorem S4 (Unidirectional or bidirectional compression without memory)**.** *Considering that Assumptions 1 to 6 hold. Taking γ such that*

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

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

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

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

Proof. In the case of variant of `Artemis` with $\alpha=0$, we don't have any memory term, thus $p=0$ and we don't need to use the Lyapunov function.

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

$$\begin{aligned}\left\|w_{k+1}-w_*\right\|^2 &= \left\|w_k - \gamma G_{k+1,S_k} - w_*\right\|^2\\ &= \left\|w_k-w_*\right\|^2 - 2\gamma\left\langle G_{k+1,S_k} \mid w_k-w_*\right\rangle + \gamma^2\left\|G_{k+1,S_k}\right\|^2\,.\end{aligned}$$

First, we have $\mathbb{E}\left[G_{k+1,S_k} \mid \sigma(\mathcal{G}_{k+1}\cup\mathcal{B}_k)\right] = \widehat{g}_{k+1,S_k}$; secondly considering that:

$$\mathbb{E}\left[\left\|G_{k+1,S_k}\right\|^2 \,\middle|\, \sigma(\mathcal{G}_{k+1}\cup\mathcal{B}_k)\right] = \mathbb{V}(G_{k+1,S_k}) + \left\|\mathbb{E}\left[G_{k+1,S_k} \mid \sigma(\mathcal{G}_{k+1}\cup\mathcal{B}_k)\right]\right\|^2 = (\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left\|\widehat{g}_{k+1,S_k}\right\|^2\,,$$

it leads to:

$$\begin{aligned}\mathbb{E}\left[\left\|w_{k+1}-w_*\right\|^2 \,\middle|\, \sigma(\mathcal{G}_{k+1}\cup\mathcal{B}_k)\right] &= \mathbb{E}\left[\left\|w_k-w_*\right\|^2 \,\middle|\, \sigma(\mathcal{G}_{k+1}\cup\mathcal{B}_k)\right] - 2\gamma\left\langle \widehat{g}_{k+1,S_k} \mid w_k-w_*\right\rangle\\ &\quad + \gamma^2(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left\|\widehat{g}_{k+1,S_k}\right\|^2\,.\end{aligned}$$

Now, if we take expectation w.r.t σ-algebra $\mathcal{I}_k \subset \sigma(\mathcal{G}_{k+1}\cup\mathcal{B}_k)$ (the inclusion is true because $\mathcal{G}_{k+1}$ contains $\mathcal{B}_{k-1}$); with use of Propositions S2, S3 and S7 and Lemma S9 we obtain :

$$\begin{aligned}\mathbb{E}\left[\left\|w_{k+1}-w_*\right\|^2 \,\middle|\, \mathcal{I}_k\right] &= \mathbb{E}\left[\left\|w_k-w_*\right\|^2 \,\middle|\, \mathcal{I}_k\right]\\ &\quad - 2\gamma\left\langle \nabla F(w_k) \mid w_k-w_*\right\rangle\\ &\quad + \gamma^2(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\frac{(1-p)}{pN^2}\sum_{i=0}^{N}\mathbb{E}\left[\left\|\widehat{g}_{k+1}^i\right\|^2 \,\middle|\, \mathcal{I}_k\right]\\ &\quad + \gamma^2(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\mathbb{E}\left[\left\|\widehat{g}_{k+1}\right\|^2 \,\middle|\, \mathcal{I}_k\right]\,.\end{aligned} \tag{S10}$$

For sake of clarity we temporarily note: $\Pi = \dfrac{(1-p)}{pN^2} \sum_{i=0}^N \mathbb{E}\left[\left\|\widehat{g}_{k+1}^i\right\|^2 \;\middle|\; \mathcal{I}_k\right] + \mathbb{E}\left[\left\|\widehat{g}_{k+1}\right\|^2 \;\middle|\; \mathcal{I}_k\right]$. Recall that in fact, Π is equal to $\mathbb{E}\left[\left\|\widehat{g}_{k+1,S_k}\right\|^2 \;\middle|\; \mathcal{I}_k\right]$.

Using Lemmas S10 and S11, we have:

$$
\begin{aligned}
\mathbb{E}\left[\Pi \mid \mathcal{I}_k\right] \leq& \frac{(1-p)(\omega_{\mathcal{C}}^{\mathrm{up}}+1)}{pN^2} \sum_{i=0}^N \mathbb{E}\left[\left\|g_{k+1}^i\right\|^2 \;\middle|\; \mathcal{I}_k\right] \\
&+ \frac{\omega_{\mathcal{C}}^{\mathrm{up}}}{N^2} \sum_{i=0}^N \mathbb{E}\left[\left\|g_{k+1}^i\right\|^2 \;\middle|\; \mathcal{I}_k\right] + \frac{1}{N} \sum_{i=0}^N \mathbb{E}\left[\left\|g_{k+1}^i - h_*^i\right\|^2 \;\middle|\; \mathcal{I}_k\right] \\
&+ L\left\langle \nabla F(w_k) \mid w_k - w_* \right\rangle . \\
\leq& \frac{(\omega_{\mathcal{C}}^{\mathrm{up}}+1-p)}{pN^2} \sum_{i=0}^N \mathbb{E}\left[\left\|g_{k+1}^i\right\|^2 \;\middle|\; \mathcal{I}_k\right] + \frac{1}{N} \sum_{i=0}^N \mathbb{E}\left[\left\|g_{k+1}^i - h_*^i\right\|^2 \;\middle|\; \mathcal{I}_k\right] \\
&+ L\left\langle \nabla F(w_k) \mid w_k - w_* \right\rangle .
\end{aligned}
$$

Lets introducing the noise at optimal point w_* with the two equations of Lemma S8:

$$
\begin{aligned}
\mathbb{E}\left[\Pi \mid \mathcal{I}_k\right] \leq& \frac{(\omega_{\mathcal{C}}^{\mathrm{up}}+1-p)}{pN^2} \sum_{i=1}^N 2\left(\mathbb{E}\left[\left\|g_{k+1}^i - g_{k+1,*}^i\right\|^2 \;\middle|\; \mathcal{I}_k\right] + (\frac{\sigma_*^2}{b} + B^2)\right) \\
&+ \frac{1}{N^2} \sum_{i=1}^N 2\left(\mathbb{E}\left[\left\|g_{k+1}^i - g_{k+1,*}^i\right\|^2 \;\middle|\; \mathcal{I}_k\right] + \frac{\sigma_*^2}{b}\right) \\
&+ L\left\langle \nabla F(w_k) \mid w_k - w_* \right\rangle . \\
\leq& \frac{2(\omega_{\mathcal{C}}^{\mathrm{up}}+1)}{pN^2} \sum_{i=1}^N \mathbb{E}\left[\left\|g_{k+1}^i - g_{k+1,*}^i\right\|^2 \;\middle|\; \mathcal{I}_k\right] \\
&+ L\left\langle \nabla F(w_k) \mid w_k - w_* \right\rangle + \frac{2 \times \left((\omega_{\mathcal{C}}^{\mathrm{up}}+1)\frac{\sigma_*^2}{b} + \omega_{\mathcal{C}}^{\mathrm{up}} + 1 - p)B^2\right)}{pN} .
\end{aligned}
$$

Invoking cocoercivity (Assumption 2):

$$
\begin{aligned}
\mathbb{E}\left[\Pi \mid \mathcal{I}_k\right] \leq& \frac{2(\omega_{\mathcal{C}}^{\mathrm{up}}+1)}{pN^2} \sum_{i=1}^N \mathbb{E}\left[L\left\langle g_{k+1}^i - g_{k+1,*}^i \;\middle|\; w_k - w_* \right\rangle \;\middle|\; \mathcal{I}_k\right] \\
&+ L\left\langle \nabla F(w_k) \mid w_k - w_* \right\rangle + \frac{2 \times \left((\omega_{\mathcal{C}}^{\mathrm{up}}+1)\frac{\sigma_*^2}{b} + \omega_{\mathcal{C}}^{\mathrm{up}} + 1 - p)B^2\right)}{pN} \\
\leq& \frac{2(\omega_{\mathcal{C}}^{\mathrm{up}}+1)L}{pN} \left\langle \nabla F(w_k) \mid w_k - w_* \right\rangle \\
&+ L\left\langle \nabla F(w_k) \mid w_k - w_* \right\rangle + \frac{2 \times \left((\omega_{\mathcal{C}}^{\mathrm{up}}+1)\frac{\sigma_*^2}{b} + \omega_{\mathcal{C}}^{\mathrm{up}} + 1 - p)B^2\right)}{pN} .
\end{aligned}
\tag{S11}
$$

Finally, we can inject eq. (S11) in eq. (S10) to obtain:

$$\mathbb{E}\left[\|w_{k+1}-w_*\|^2 \;\middle|\; \mathcal{I}_k\right] \leq \|w_k-w_*\|^2 - 2\gamma\left(1-\frac{\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)(\omega_{\mathcal{C}}^{\mathrm{up}}+1)}{pN} - \frac{\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)}{2}\right)\langle \nabla F(w_k) \mid w_k - w_*\rangle$$
$$+ \frac{\overbrace{2\times(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left((\omega_{\mathcal{C}}^{\mathrm{up}}+1)\frac{\sigma_*^2}{b} + (\omega_{\mathcal{C}}^{\mathrm{up}}+1-p)B^2\right)}^{=E}}{pN}\,. \tag{S12}$$

We need $1-\frac{\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)(\omega_{\mathcal{C}}^{\mathrm{up}}+1)}{pN} - \frac{\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)}{2} \geq 0$ in order to further apply strong convexity. This condition is equivalent to:

$$\gamma \leq \frac{2pN}{L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left(pN+2(\omega_{\mathcal{C}}^{\mathrm{up}}+1)\right)}\,.$$

Finally, using strong convexity of F (Assumption 1), we rewrite Equation (S12):

$$\mathbb{E}\left[\|w_{k+1}-w_*\|^2 \;\middle|\; \mathcal{I}_k\right] \leq \|w_k-w_*\|^2 - 2\gamma\mu\left(1-\frac{\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)(\omega_{\mathcal{C}}^{\mathrm{up}}+1)}{pN} - \frac{\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)}{2}\right)\|w_k-w_*\|^2$$
$$+ 2\gamma^2\frac{E}{pN}\,, \text{ equivalent to:}$$
$$\leq \left(1-2\gamma\mu\left(1-\frac{\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)(\omega_{\mathcal{C}}^{\mathrm{up}}+1)}{pN} - \frac{\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)s}{2}\right)\right)\|w_k-w_*\|^2 + 2\gamma^2\frac{E}{pN}\,.$$

To guarantee a convergence in $(1-\gamma\mu)$, we need:

$$\begin{aligned} & \frac{1}{2} \geq \frac{\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)(\omega_{\mathcal{C}}^{\mathrm{up}}+1)}{pN} + \frac{\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)}{2} \\ \iff \quad & \gamma \leq \frac{pN}{L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left(pN+2(\omega_{\mathcal{C}}^{\mathrm{up}}+1)\right)}\,, \end{aligned}$$

which is stronger than the condition obtained to correctly apply strong convexity. Then we are allowed to write:

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

and the proof is complete.

□

E.2 Proof of Theorem S5 - variant with memory of Artemis

Theorem S5 (Unidirectional or bidirectional compression with memory)**.** *Considering that Assumptions 1 to 6 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:*

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

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

$$\frac{\omega_{\mathcal{C}}^{\mathrm{up}}(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)}{\alpha\left(3-2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}}+1)\right)} \leq C \leq \frac{Np-\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left(Np+2(2\omega_{\mathcal{C}}^{\mathrm{up}}+1)\right)}{4\gamma L\alpha\left(2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}}+1)-1\right)}\,. \tag{S13}$$

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

$$\begin{cases} \dfrac{1}{2(\omega_{\mathcal{C}}^{\mathrm{up}}+1)} \leq \alpha < \min\left(\dfrac{3}{2(\omega_{\mathcal{C}}^{\mathrm{up}}+1)},\quad \dfrac{3Np-\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left(3Np+8\omega_{\mathcal{C}}^{\mathrm{up}}+6\right)}{2(\omega_{\mathcal{C}}^{\mathrm{up}}+1)(Np-\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)(Np+2))}\right) \\ \gamma < \min\left\{\begin{array}{ll} \dfrac{1}{(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left(1+\dfrac{2}{Np}\right)L}, & \dfrac{3}{(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left(3+\dfrac{6+8\omega_{\mathcal{C}}^{\mathrm{up}}}{Np}\right)L}, \\ \dfrac{Np}{(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left(Np+2(2\omega_{\mathcal{C}}^{\mathrm{up}}+1)\right)L} & \end{array}\right\} \end{cases}. \tag{S14}$$

And we have a bound for the Lyapunov function:

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

with

$$E=\frac{\sigma_*^2}{b}\left((2\omega_{\mathcal{C}}^{\mathrm{up}}+1)(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)+4\alpha^2C(\omega_{\mathcal{C}}^{\mathrm{up}}+1)-2\alpha C\right)\,.$$

In the case of unidirectional compression (resp. no compression), we have $\omega_{\mathcal{C}}^{\mathrm{dwn}}=0$ (resp. $\omega_{\mathcal{C}}^{\mathrm{up/dwn}}=0$).

Proof. Let $k \in [\![1,K]\!]$, by definition of the update schema:

$$\begin{aligned} \left\|w_{k+1}-w_*\right\|^2 &= \left\|w_k-w_*+\gamma G_{k+1,S_k}\right\|^2 \\ &= \left\|w_k-w_*\right\|-2\gamma\left\langle G_{k+1,S_k}\mid w_k-w_*\right\rangle+\gamma^2\left\|G_{k+1,S_k}\right\|^2\,. \end{aligned}$$

First, we have $\mathbb{E}\left[G_{k+1,S_k}\mid\sigma(\mathcal{G}_{k+1}\cup\mathcal{B}_k)\right]=\widehat{g}_{k+1,S_k}$; secondly considering that:

$$\mathbb{E}\left[\left\|G_{k+1,S_k}\right\|^2\,\middle|\,\sigma(\mathcal{G}_{k+1}\cup\mathcal{B}_k)\right]=\mathbb{V}(G_{k+1,S_k})+\left\|\mathbb{E}\left[G_{k+1,S_k}\mid\sigma(\mathcal{G}_{k+1}\cup\mathcal{B}_k)\right]\right\|^2=(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left\|\widehat{g}_{k+1,S_k}\right\|^2\,,$$

it leads to:

$$\begin{aligned} \mathbb{E}\left[\left\|w_{k+1}-w_*\right\|^2\,\middle|\,\sigma(\mathcal{G}_{k+1}\cup\mathcal{B}_k)\right] &= \mathbb{E}\left[\left\|w_k-w_*\right\|^2\,\middle|\,\sigma(\mathcal{G}_{k+1}\cup\mathcal{B}_k)\right]-2\gamma\left\langle\widehat{g}_{k+1,S_k}\mid w_k-w_*\right\rangle \\ &\quad+\gamma^2(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left\|\widehat{g}_{k+1,S_k}\right\|^2\,. \end{aligned}$$

we can invoke Lemma S13 with the σ_*-algebras $\mathcal{I}_k$:

$$\begin{aligned} \mathbb{E}\left[\left\|w_{k+1}-w_*\right\|^2\,\middle|\,\mathcal{I}_k\right] \leq \left\|w_k-w_*\right\|^2 &-2\gamma\mathbb{E}\left[\left\langle\widehat{g}_{k+1,S_k}\mid w_k-w_*\right\rangle\;\middle|\;\mathcal{I}_k\right] \\ &+\frac{2(2\omega_{\mathcal{C}}^{\mathrm{up}}+1)(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\gamma^2}{pN^2}\sum_{i=1}^N\mathbb{E}\left[\left\|g_{k+1}^i-g_{k+1,*}^i\right\|^2\,\middle|\,\mathcal{I}_k\right] \\ &+\frac{2\omega_{\mathcal{C}}^{\mathrm{up}}(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\gamma^2}{pN^2}\sum_{i=1}^N\mathbb{E}\left[\left\|h_k^i-h_*^i\right\|^2\,\middle|\,\mathcal{I}_k\right] \\ &+\gamma^2(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)L\left\langle\nabla F(w_k)\mid w_k-w_*\right\rangle \\ &+\frac{2(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)(2\omega_{\mathcal{C}}^{\mathrm{up}}+1)\gamma^2}{pN^2}\times\frac{\sigma_*^2}{b}\,. \end{aligned} \tag{S15}$$

Note that in the case of unidirectional compression, we have $G_{k+1} = \widehat{g}_{k+1}$, and the steps above are more straightforward. Recall that according to Lemma S14 (and taking the sum), we have:

$$
\begin{aligned}
&\frac{1}{p^2N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|h_{k+1}^i - h_*^i\right\|^2 \ \middle|\ \mathcal{I}_k\right] \\
&\qquad\leq \left(1 + p(2\alpha^2\omega_{\mathcal{C}}^{\mathrm{up}} + 2\alpha^2 - 3\alpha)\right)\frac{1}{p^2N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \ \middle|\ \mathcal{I}_k\right] \\
&\qquad+ 2(2\alpha^2\omega_{\mathcal{C}}^{\mathrm{up}} + 2\alpha^2 - \alpha)\frac{1}{pN^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_{k+1}^i - g_{k+1,*}^i\right\|^2 \ \middle|\ \mathcal{I}_k\right] \\
&\qquad+ \frac{2}{pN}\frac{\sigma_*^2}{b}\left(2\alpha^2(\omega_{\mathcal{C}}^{\mathrm{up}} + 1) - \alpha\right)
\end{aligned}
\tag{S16}
$$

With a linear combination (S15) + $2\gamma^2 C$ (S16):

$$
\begin{aligned}
&\mathbb{E}\left[\left\|w_{k+1} - w_*\right\|^2 \ \middle|\ \mathcal{I}_k\right] + 2\gamma^2 C\frac{1}{p^2N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|h_{k+1}^i - h_*^i\right\|^2 \ \middle|\ \mathcal{I}_k\right] \\
&\quad\leq \left\|w_k - w_*\right\|^2 - 2\gamma\mathbb{E}\left[\langle \widehat{g}_{k+1,S_k} \mid w_k - w_*\rangle \ \mid \mathcal{I}_k\right] \\
&\qquad + 2\gamma^2\left((2\omega_{\mathcal{C}}^{\mathrm{up}} + 1)(\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1) + 4\alpha^2 C\omega_{\mathcal{C}}^{\mathrm{up}} + 4\alpha^2 C - 2\alpha C\right) \\
&\qquad\qquad \times \frac{1}{pN^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_{k+1}^i - g_{k+1,*}^i\right\| \ \mid \mathcal{I}_k\right] \\
&\qquad + 2\gamma^2 C\underbrace{\left(\frac{p\omega_{\mathcal{C}}^{\mathrm{up}}(\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1)}{C} + 1 + p(2\alpha^2\omega_{\mathcal{C}}^{\mathrm{up}} + 2\alpha^2 - 3\alpha)\right)}_{:=D_C}\frac{1}{p^2N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \ \middle|\ \mathcal{I}_k\right] \\
&\qquad + \gamma^2(\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1)L\left\langle \nabla F(w_k) \mid w_k - w_*\right\rangle \\
&\qquad + \frac{2\gamma^2}{pN}\underbrace{\left(\frac{\sigma_*^2}{b}\left((2\omega_{\mathcal{C}}^{\mathrm{up}} + 1)(\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1) + 4\alpha^2 C(\omega_{\mathcal{C}}^{\mathrm{up}} + 1) - 2\alpha C\right)\right)}_{:=E}.
\end{aligned}
$$

Because $\mathcal{H}_k \subset \mathcal{F}_{k+1}$, because $\widehat{g}_{k+1}$ is independent of $\mathcal{B}_k$, and with Propositions S2 and S3:

$$
\mathbb{E}\left[\widehat{g}_{k+1} \mid \mathcal{I}_k\right] = \mathbb{E}\left[\mathbb{E}\left[\widehat{g}_{k+1} \mid \mathcal{F}_{k+1}\right] \ \mid \mathcal{H}_k\right] = \mathbb{E}\left[g_{k+1} \mid \mathcal{H}_k\right] = \nabla F(w_k)\,.
$$

We transform $\left\|g_{k+1}^i - g_{k+1,*}^i\right\|$ applying co-coercivity (Lemma S15):

$$
\begin{aligned}
\mathbb{E}V_{k+1} &\leq \left\|w_k - w_*\right\|^2 \\
&\qquad - 2\gamma\left(1 - \gamma L\left(\frac{\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1}{2} + \frac{\overbrace{(2\omega_{\mathcal{C}}^{\mathrm{up}} + 1)(\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1) + 4\alpha^2 C\omega_{\mathcal{C}}^{\mathrm{up}} + 4\alpha^2 C - 2\alpha C}^{:=A_C}}{pN}\right)\right) \\
&\qquad\qquad \times \left\langle \nabla F(w_k) \mid w_k - w_*\right\rangle \\
&\qquad + 2\gamma^2 C D_C\frac{1}{p^2N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \ \middle|\ \mathcal{I}_k\right] \\
&\qquad + \frac{2\gamma^2}{pN}E\,.
\end{aligned}
\tag{S17}
$$

Now, the goal is to apply strong convexity of F (Assumption 1) using the inequality presented in Lemma S4. But then we must have:

$$1 - \gamma L \left(\frac{\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1}{2} + \frac{A_C}{pN} \right) \geq 0 \,,$$

However, in order to later obtain a convergence in $(1-\gamma\mu)$, we will use a stronger condition and, instead, state that we need:

$$\begin{aligned}
& \gamma L \left(\frac{\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1}{2} + \frac{A_C}{pN} \right) \leq \frac{1}{2} \\
\Longleftrightarrow \quad & A_C \leq \frac{\left(1 - \gamma L (\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1)\right) Np}{2\gamma L} \\
\Longleftrightarrow \quad & 2\alpha C \left(2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}} + 1) - 1\right) + (2\omega_{\mathcal{C}}^{\mathrm{up}} + 1)(\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1) \leq \frac{(1 - \gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1))Np}{2\gamma L} \\
\Longleftrightarrow \quad & C \leq \frac{Np - \gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1)\left(Np + 2(2\omega_{\mathcal{C}}^{\mathrm{up}} + 1)\right)}{4\gamma L \alpha \left(2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}} + 1) - 1\right)} \,.
\end{aligned}$$

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

$$\begin{cases}
Np - \gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1)\left(Np + 2(2\omega_{\mathcal{C}}^{\mathrm{up}} + 1)\right) > 0 \Longleftrightarrow \gamma < \dfrac{Np}{(\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1)\left(Np + 2(2\omega_{\mathcal{C}}^{\mathrm{up}} + 1)\right) L} \\
2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}} + 1) - 1 \leq 0 \Longleftrightarrow \alpha \geq \dfrac{1}{2(\omega_{\mathcal{C}}^{\mathrm{up}} + 1)} \,.
\end{cases}$$

Strong convexity is applied, and we obtain:

$$\begin{aligned}
\mathbb{E} V_{k+1} \leq & \left(1 - 2\gamma\mu\left(1 - \frac{\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1)}{2} - \frac{\gamma L A_C}{pN}\right)\right) \|w_k - w_*\|^2 \\
& + 2\gamma^2 C D_C \frac{1}{p^2 N} \sum_{i=1}^N \mathbb{E}\left[\left\| h_k^i - h_*^i \right\|^2 \,\middle|\, \mathcal{I}_k \right] \\
& + 2\gamma^2 \frac{E}{pN} \,.
\end{aligned} \tag{S18}$$

To guarantee a $(1-\gamma\mu)$ convergence, constants must verify:

$$\begin{cases}
\dfrac{\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1)}{2} - \dfrac{\gamma L A_C}{pN} \leq \frac{1}{2} \quad \text{(which is already verified)} \\
D_C \leq 1 - \gamma\mu \Longleftrightarrow \dfrac{\omega_{\mathcal{C}}^{\mathrm{up}}(p\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1)}{C} \leq p(3\alpha - 2\alpha^2 \omega_{\mathcal{C}}^{\mathrm{up}} - 2\alpha) - \gamma\mu \\
\qquad\qquad\qquad \Longleftrightarrow C \geq \dfrac{p\omega_{\mathcal{C}}^{\mathrm{up}}(\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1)}{\alpha p \left(3 - 2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}} + 1)\right) - \gamma\mu}
\end{cases} .$$

In the following we will consider that $\frac{\gamma\mu}{\alpha} = \underset{\mu \to 0}{o}(1)$ which is possible because α is independent of μ (it depends only of $\omega_{\mathcal{C}}^{\mathrm{up}}$ and $\omega_{\mathcal{C}}^{\mathrm{dwn}}$) and it result to:

$$\alpha p \left(3 - 2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}} + 1)\right) - \gamma\mu \underset{\mu \to 0}{\sim} \alpha p \left(3 - 2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}} + 1)\right)$$

Thus, the condition on C becomes:

$$\frac{\omega_{\mathcal{C}}^{\mathrm{up}}(\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1)}{\alpha \left(3 - 2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}} + 1)\right)} \leq C \,,$$

which is correct only if $\alpha \leq \dfrac{3}{2(\omega_{\mathcal{C}}^{\mathrm{up}}+1)}$.

And we obtain the following conditions on C:

$$\frac{\omega_{\mathcal{C}}^{\mathrm{up}}(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)}{\alpha\left(3-2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}}+1)\right)} \leq C \leq \frac{Np-\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left(Np+2(2\omega_{\mathcal{C}}^{\mathrm{up}}+1)\right)}{4\gamma L\alpha\left(2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}}+1)-1\right)}\,.$$

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

$$\frac{\omega_{\mathcal{C}}^{\mathrm{up}}(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)}{\alpha\left(3-2\alpha(2\omega_{\mathcal{C}}^{\mathrm{up}}+1)\right)} \leq \frac{Np-\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left(Np+2(\omega_{\mathcal{C}}^{\mathrm{up}}+1)\right)}{4\gamma L\alpha\left(2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}}+1)-1\right)}$$

For sake of clarity we denote momentarily $\tilde{\gamma}=(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\gamma L$, hence the below condition becomes:

$$\begin{aligned}
& 8\alpha\omega_{\mathcal{C}}^{\mathrm{up}}(\omega_{\mathcal{C}}^{\mathrm{up}}+1)\tilde{\gamma}-4\omega_{\mathcal{C}}^{\mathrm{up}}\tilde{\gamma} \leq 3Np-3\tilde{\gamma}\left(Np+2(2\omega_{\mathcal{C}}^{\mathrm{up}}+1)\right)\\
&\qquad\qquad -2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}}+1)Np+2\alpha\tilde{\gamma}(\omega_{\mathcal{C}}^{\mathrm{up}}+1)\left(Np+2(2\omega_{\mathcal{C}}^{\mathrm{up}}+1)\right)\\
\iff\quad & 2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}}+1)(4\omega_{\mathcal{C}}^{\mathrm{up}}-Np-4\omega_{\mathcal{C}}^{\mathrm{up}}-2)\tilde{\gamma}+2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}}+1)Np\\
&\quad \leq 3Np-\tilde{\gamma}\left(3Np+6(2\omega_{\mathcal{C}}^{\mathrm{up}}+1)-4\omega_{\mathcal{C}}^{\mathrm{up}}\right)\\
\iff\quad & 2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}}+1)(Np-\tilde{\gamma}(N+2)) \leq 3Np-\tilde{\gamma}\left(3Np+8\omega_{\mathcal{C}}^{\mathrm{up}}+6\right)
\end{aligned}$$

And at the end, we obtain:

$$\alpha \leq \frac{3Np-\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left(3Np+8\omega_{\mathcal{C}}^{\mathrm{up}}+6\right)}{2(\omega_{\mathcal{C}}^{\mathrm{up}}+1)(Np-\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)(Np+2))}\,.$$

Again, this implies two conditions on gamma:

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

The constant C exists, and from eq. (S18) we are allowed to write:

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

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

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

Thus, we highlighted that the Lyapunov function

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

is a $(1-\gamma\mu)$ contraction if C is taken in a given interval, with γ and α satisfying some conditions. This guarantee the convergence of the Artemis using version 1 or 2 with $\alpha\neq 0$ (algorithm with uni-compression or bi-compression combined with a memory mechanism).

□

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

Theorem S6 (Unidirectional or bidirectional compression using memory and averaging)**.** *We suppose now that F is convex, thus $\mu = 0$ and we consider that Assumptions 2 to 6 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:*

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

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

$$\frac{\omega_{\mathcal{C}}^{\mathrm{up}}(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)}{\alpha\left(3-2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}}+1)\right)} \leq C \leq \frac{Np - \gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left(Np+2(2\omega_{\mathcal{C}}^{\mathrm{up}}+1)\right)}{4\gamma L\alpha\left(2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}}+1)-1\right)} . \tag{S19}$$

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

$$\begin{cases} \dfrac{1}{2(\omega_{\mathcal{C}}^{\mathrm{up}}+1)} \leq \alpha < \min\left(\dfrac{3}{2(\omega_{\mathcal{C}}^{\mathrm{up}}+1)}, \quad \dfrac{3Np - \gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left(3Np+8\omega_{\mathcal{C}}^{\mathrm{up}}+6\right)}{2(\omega_{\mathcal{C}}^{\mathrm{up}}+1)(Np-\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)(Np+2))}\right) \\ \gamma < \min\left\{ \begin{array}{l} \dfrac{1}{(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left(1+\dfrac{2}{Np}\right)L}, \quad \dfrac{3}{(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left(3+\dfrac{6+8\omega_{\mathcal{C}}^{\mathrm{up}}}{Np}\right)L}, \\ \dfrac{Np}{(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left(Np+2(2\omega_{\mathcal{C}}^{\mathrm{up}}+1)\right)L} \end{array} \right\} \end{cases} . \tag{S20}$$

And we have the following bound:

$$F\left(\frac{1}{K}\sum_{k=0}^K w_k\right) - F(w_*) \leq \frac{\|w_0 - w_*\|^2 + 2C\dfrac{\gamma^2 B^2}{p^2}}{\gamma K} + 2\gamma \frac{E}{pN} , \tag{S21}$$

with $E = \frac{\sigma_*^2}{b}\left((2\omega_{\mathcal{C}}^{\mathrm{up}}+1)(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1) + 4\alpha^2 C(\omega_{\mathcal{C}}^{\mathrm{up}}+1) - 2\alpha p\right)$.

Equation (S21) can be written as in Theorem 2 if we take $\gamma = \min\left(\sqrt{\frac{pN\delta_0^2}{2EK}}; \gamma_{\max}\right)$, where $\gamma_{\max}$ is the maximal possible value of γ as precised by Equation (S20):

$$F\left(\frac{1}{K}\sum_{k=0}^K w_k\right) - F(w_*) \leq 2\max\left(\sqrt{\frac{2\delta_0^2 E}{pNK}}; \frac{\delta_0^2}{\gamma_{\max} K}\right) + \frac{2\gamma_{\max} C B^2}{p^2 K}$$

Proof. Starting from eq. (S17) from the proof of Theorem S5:

$$\begin{aligned} \mathbb{E} V_{k+1} \leq\ & \|w_k - w_*\|^2 \\ & - 2\gamma\left(1 - \gamma L\left(\frac{\omega_{\mathcal{C}}^{\mathrm{dwn}}+1}{2} + \frac{\overbrace{(2\omega_{\mathcal{C}}^{\mathrm{up}}+1)(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1) + 4\alpha^2 C\omega_{\mathcal{C}}^{\mathrm{up}} + 4\alpha^2 C - 2\alpha C}^{:=A_C}}{pN}\right)\right) \\ & \qquad \times \langle \nabla F(w_k) \mid w_k - w_* \rangle \\ & + 2\gamma^2 C D_C \frac{1}{p^2 N^2} \sum_{i=1}^N \mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \,\middle|\, \mathcal{I}_k\right] \\ & + \frac{2\gamma^2}{pN} \underbrace{\left(\frac{\sigma_*^2}{b}\left((2\omega_{\mathcal{C}}^{\mathrm{up}}+1)(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1) + 4\alpha^2 C(\omega_{\mathcal{C}}^{\mathrm{up}}+1) - 2\alpha p\right)\right)}_{:=E} . \end{aligned}$$

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

$$\begin{aligned}\mathbb{E}V_{k+1} \leq \|w_k - w_*\|^2 \\ &- 2\gamma \left(1 - \gamma L \left(\frac{\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1}{2} + \frac{A_C}{pN}\right)\right)(F(w_k) - F(w_*)) \\ &+ 2\gamma^2 C D_C \frac{1}{p^2N^2} \sum_{i=1}^N \mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{I}_k\right] \\ &+ \frac{2\gamma^2}{pN} E\,.\end{aligned} \tag{S22}$$

As in Theorem S5, we want:

$$\begin{aligned} &\gamma L \left(\frac{\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1}{2} + \frac{A_C}{pN}\right) \leq \frac{1}{2} \\ \iff \qquad & C \leq \frac{Np - \gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1)\left(Np + 2(2\omega_{\mathcal{C}}^{\mathrm{up}} + 1)\right)}{4\gamma L \alpha \left(2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}} + 1) - 1\right)}\,, \end{aligned} \tag{S23}$$

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

$$\begin{cases} pN - \gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1)\left(pN + 2(2\omega_{\mathcal{C}}^{\mathrm{up}} + 1)\right) \geq 0 \iff \gamma \leq \dfrac{pN}{(\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1)\left(Np + 2(2\omega_{\mathcal{C}}^{\mathrm{up}} + 1)\right) L} \\ 2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}} + 1) - 1 \leq 0 \iff \alpha \geq \dfrac{1}{2(\omega_{\mathcal{C}}^{\mathrm{up}} + 1)}\,. \end{cases}$$

Returning to eq. (S22), taking benefit of eq. (S23) and passing $F(w_k) - F(w_*)$ on the left side gives:

$$\gamma(F(w_k) - F(w_*)) \leq \|w_k - w_*\|^2 + 2\gamma^2 C D_C \frac{1}{p^2N^2} \sum_{i=1}^N \mathbb{E}\left[\left\|h_k^i - h_*^i\right\|^2 \;\middle|\; \mathcal{I}_k\right] - \mathbb{E}V_{k+1} + \frac{2\gamma^2}{pN} E\,.$$

If $D_C \leq 1$, then:

$$\gamma(F(w_k) - F(w_*)) \leq \mathbb{E}V_k - \mathbb{E}V_{k+1} + \frac{2\gamma^2}{pN} E\,,$$

summing over all K iterations:

$$\begin{aligned}\gamma\left(\frac{1}{K}\sum_{k=0}^K F(w_k) - F(w_*)\right) &\leq \frac{1}{K}\sum_{k=0}^K \left(\mathbb{E}V_k - \mathbb{E}V_{k+1} + 2\gamma^2 \frac{E}{pN}\right) \\ &\leq \frac{\mathbb{E}V_0 - \mathbb{E}V_{K+1}}{K} + 2\gamma^2 \frac{E}{pN} \quad \text{because } E \text{ is independent of } K.\end{aligned}$$

Thus, by convexity:

$$F\left(\frac{1}{K}\sum_{k=0}^K w_k\right) - F(w_*) \leq \frac{1}{K}\sum_{k=0}^K F(w_k) - F(w_*) \leq \frac{V_0}{\gamma K} + 2\gamma \frac{E}{pN}\,.$$

Last step is to extract conditions over γ and α from requirement $D_C \leq 1$:

$$D_C < 1 \iff \frac{\omega_{\mathcal{C}}^{\mathrm{up}}(\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1)}{C} < 3\alpha - 2\alpha^2 \omega_{\mathcal{C}}^{\mathrm{up}} - 2\alpha \iff p > \frac{\omega_{\mathcal{C}}^{\mathrm{up}}(\omega_{\mathcal{C}}^{\mathrm{dwn}} + 1)}{\alpha\left(3 - 2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}} + 1)\right)}\,,$$

and the second inequality is correct only if $\alpha \leq \dfrac{3}{2(\omega_{\mathcal{C}}^{\mathrm{up}} + 1)}$.

From this development follows the following conditions on p, which are equivalent to those obtain in Theorem S5

$$\frac{\omega_{\mathcal{C}}^{\mathrm{up}}(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)}{\alpha\left(3-2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}}+1)\right)} \leq C \leq \frac{Np-\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left(Np+2(2\omega_{\mathcal{C}}^{\mathrm{up}}+1)\right)}{4\gamma L\alpha\left(2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}}+1)-1\right)}\,.$$

This interval is not empty:

$$\begin{aligned} & \frac{\omega_{\mathcal{C}}^{\mathrm{up}}(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)}{\alpha\left(3-2\alpha(2\omega_{\mathcal{C}}^{\mathrm{up}}+1)\right)} < \frac{N-\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left(N+2(\omega_{\mathcal{C}}^{\mathrm{up}}+1)\right)}{4\gamma L\alpha\left(2\alpha(\omega_{\mathcal{C}}^{\mathrm{up}}+1)-1\right)} \\ \iff \qquad & \alpha < \frac{3N-\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)\left(3N+8\omega_{\mathcal{C}}^{\mathrm{up}}+6\right)}{2(\omega_{\mathcal{C}}^{\mathrm{up}}+1)(N-\gamma L(\omega_{\mathcal{C}}^{\mathrm{dwn}}+1)(N+2))}\,. \end{aligned}$$

Again, this implies two conditions on gamma:

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

which guarantees the existence of p and thus the validity of the above development.

As a conclusion:

$$\begin{aligned} F\left(\frac{1}{K}\sum_{k=0}^{K} w_k\right) - F(w_*) &\leq \frac{V_0}{\gamma K} + 2\gamma\frac{E}{pN} \leq \frac{\left\|w_0-w_*\right\|^2 + 2C\dfrac{\gamma^2 B^2}{p^2}}{\gamma K} + 2\gamma\frac{E}{pN}\,. \\ &\leq \frac{\left\|w_0-w_*\right\|^2}{\gamma K} + 2\gamma\left(\frac{E}{pN}+\frac{CB^2}{p^2K}\right)\,. \end{aligned}$$

Next, our goal is to define the optimal step size γ_{opt}. With this aim, we bound $2\gamma\frac{CB^2}{p^2K}$ by $2\gamma_{\max}\frac{CB^2}{p^2K}$. 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 $\frac{B^2}{K}$ is "small". And we obtain:

$$F\left(\frac{1}{K}\sum_{k=0}^{K} w_k\right) - F(w_*) \leq \frac{\left\|w_0-w_*\right\|^2}{\gamma K} + 2\gamma\frac{E}{pN} + 2\gamma_{\max}\frac{CB^2}{p^2K}.$$

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

$$\gamma_{opt} = \sqrt{\frac{\left\|w_0-w_*\right\|^2 pN}{2EK}}\,,$$

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

$$F\left(\frac{1}{K}\sum_{k=0}^{K} w_k\right) - F(w_*) \leq 2\max\left(\sqrt{\frac{2\left\|w_0-w_*\right\|^2 E}{pNK}}; \frac{\left\|w_0-w_*\right\|^2}{\gamma_{\max}K}\right) + \frac{2\gamma_{max}CB^2}{p^2K}\,.$$

□

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 S9 and S10, given in Appendices E.4.2 and E.4.3. We first introduce a few notations in Appendix E.4.1.

We consider in this section the *Stochastic Sparsification* compression operator $\mathcal{C}_q$, which is defined as follows: for any $x \in \mathbb{R}^d$, $\mathcal{C}_q(x) \overset{dist}{=} \frac{1}{q}(x_1 B_1, \dots, x_d B_d)$, with $(B_1, \dots, B_d) \sim \mathcal{B}(q)^{\otimes n}$ i.i.d. Bernoullis with mean q. That is, each coordinate is independently assigned to 0 with probability $1-q$ or rescaled by a factor q^{-1} in order to get an unbiased operator.

Lemma S16. *This compression operator satisfies Assumption 5 with $\omega_{\mathcal{C}} = q^{-1} - 1$.*

Moreover, if I consider a random variable $(B_1, \dots, B_d) \sim \mathcal{B}(q)^{\otimes n}$ and define almost surely $\mathcal{C}_q(x) \overset{a.s.}{=} \frac{1}{q}(x_1 B_1, \dots, x_d B_d)$*, then we also have that for any $x, y \in \mathbb{R}^d$, $\mathcal{C}_q(x) - \mathcal{C}_q(y) = \mathcal{C}_q(x-y)$.*

E.4.1 Background on distributions and Markov Chains

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

$$\begin{cases} \qquad\qquad w_{k+1} &= w_k - \gamma \mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{pN} \sum_{i \in S_k} \mathcal{C}_{\mathrm{up}} \left(g_{k+1}^i - h_k^i \right) + h_k^i \right) \\ \forall i \in [\![1,N]\!], \quad h_{k+1}^i &= h_k^i + \alpha \mathcal{C}_{\mathrm{up}} \left(g_{k+1}^i - h_k^i \right) \end{cases} \tag{S24}$$

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 $(w_k, \tau(h_k^i)_{i \in [\![1,N]\!]})_{k \geq 0}$ 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 and Tweedie [2009] provide an introduction to Markov chain theory. For readability, we now denote $(h_k^i)_i$ for $(h_k^i)_{i \in [\![1,N]\!]}$.

Definition 3. *For any initial distribution ν_0 on $\mathcal{B}(\mathbb{R}^{d(1+N)})$ and $k \in \mathbb{N}$, $\nu_0 R_{\gamma,v}^k$ denotes the distribution of $(w_k, \tau(h_k^i)_i)$ starting at $(w_0, \tau(h_0^i)_i)$ distributed according to ν_0.*

We can make the following comments:

1. **Initial distribution.** We consider deterministic initial points, i.e., $(w_0, \tau(h_0^i)_i)$ follows a Dirac at point $(w_0, \tau(h_0^i)_i)$. We denote this Dirac $\delta_{w_0} \otimes \otimes_{i=1}^N \delta_{\tau h_0^i} \overset{\text{not}}{=} \delta_{w_0} \otimes \delta_{\tau h_0^1} \otimes \dots \otimes \delta_{\tau h_0^N}$.
2. **Notation in the main text:** In the main text, for simplicity, we used Θ_k to denote the distribution of w_k when launched from $(w_0, \tau(h_0^i)_i)$. Thus Θ_k corresponds to the distribution of the projection on first d coordinates of $((\delta_{w_0} \otimes \otimes_{i=1}^N \delta_{\tau h_0^i}) R_\gamma^k)$.
3. **Case without memory:** In the memory-less case, we have $(h_k^i)_{k \in \mathbb{N}} \equiv 0$, 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 $(w_k, (h_k^i)_i)_{k \geq 0}$ admits a limit stationary distribution

$$\Pi_{\gamma,v} = \pi_{\gamma,v,w} \otimes \pi_{\gamma,v,(h)} \tag{S25}$$

and quantify the convergence of $((\delta_{w_0} \otimes \otimes_{i=1}^N \delta_{\tau h_0^i}) R_\gamma^k)_{k \geq 0}$ 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*

$$\mathcal{W}_2^2(\lambda, \nu) := \inf_{\xi \in \Gamma(\lambda,\nu)} \int \|x-y\|^2 \xi(dx, dy), \tag{S26}$$

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

E.4.2 Proof of the first point in Theorem 3

We prove the following proposition:

Proposition S9. *Under Assumptions 1 to 5, for $\mathcal{C}_p$ the* Stochastic Sparsification *compression operator, 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:*

$$\mathcal{W}_2^2((\delta_{w_0} \otimes \otimes_{i=1}^N \delta_{\tau h_0^i}) R_\gamma^k, \Pi_{\gamma,v}) \leq (1-\gamma\mu)^k \int_{(w',h') \in \mathbb{R}^{d(1+N)}} \left\| (w_0, \tau(h_0^i)_i) - (w', \tau(h^i)'_i) \right\|^2 \mathrm{d}\Pi_{\gamma,v}(w', (h^i)'_i).$$

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 (S25), the limit distribution of the main iterates $(w_k)_{k\in\mathbb{N}}$ and the observation that:

$$\begin{aligned}
\mathcal{W}_2^2(\Theta_k, \pi_{\gamma,v}) &\leq \mathcal{W}_2^2((\delta_{w_0} \otimes \otimes_{i=1}^N \delta_{\tau h_0^i}) R_{\gamma,v}^k, \Pi_{\gamma,v}) \\
&\leq (1-\gamma\mu)^k \int_{(w',h') \in \mathbb{R}^{d(1+N)}} \left\| (w_0, \tau(h_0^i)_i) - (w', \tau(h^i)'_i) \right\|^2 \mathrm{d}\Pi_{\gamma,v}(w', (h^i)'_i) \\
&= (1-\gamma\mu)^k C_0.
\end{aligned}$$

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:
$$\mathcal{W}_2^2(\nu_0^a R_{\gamma,v}^{k+1}, \nu_0^b R_{\gamma,v}^{k+1}) \leq (1-\gamma\mu) \mathcal{W}_2^2(\nu_0^a R_{\gamma,v}^k, \nu_0^b R_{\gamma,v}^k).$$
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 $g_{k+1}^{i,a} = g_{k+1}^{i,b}$ 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 $\mathcal{C}_{\mathrm{dwn},k}$ and $\mathcal{C}_{\mathrm{up},k}$ the compression operators at iteration k for respectively the uplink compression and downlink compression.

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

$$\begin{cases} \qquad\qquad w_{k+1}^u &= w_k^u - \gamma \mathcal{C}_{\mathrm{dwn},k}\left(\frac{1}{pN}\sum_{i\in S_k} \mathcal{C}_{\mathrm{up},k}\left(g_{k+1}^i - h_k^{i,u}\right) + h_k^{i,u}\right) \\ \forall i \in [\![1;n]\!] \quad h_{k+1}^{i,u} &= h_k^{i,u} + \alpha \mathcal{C}_{\mathrm{up},k}\left(g_{k+1}^i - h_k^{i,u}\right) \end{cases} \tag{S27}$$

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.,

$$\mathcal{W}_2^2(\nu_0^a R_{\gamma,v}^k, \nu_0^b R_{\gamma,v}^k) = \mathbb{E}\left[\left\| (w_k^a, \tau(h_k^{i,a})_i) - (w_k^b, \tau(h_k^{i,b})_i) \right\|^2\right]. \tag{S28}$$

Existence of such a couple is given by [Villani, 2009, theorem 4.1].

Then $\left((w_{k+1}^a, \tau(h_{k+1}^{i,a})_i), (w_{k+1}^b, \tau(h_{k+1}^{i,b})_i)\right)$ obtained after one update from Equation (S27) belongs to $\Gamma(\nu_0^a R_{\gamma,v}^{k+1}, \nu_0^b R_{\gamma,v}^{k+1})$, and as a consequence:

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

with $\tau^2 = 2\gamma^2 \frac{C}{p^2 N}$, where C depends on the variant as in Theorem 1.

We now follow the proof of the previous theorems to control respectively $\mathbb{E}\left[\left\|w_{k+1}^a - w_{k+1}^b\right\|^2\right]$ and $\mathbb{E}\left[\left\|h_{k+1}^{i,a} - h_{k+1}^{i,b}\right\|^2\right]$.

First, following the proof of Equation (S15), we get, using the fact that the compression operator is random sparsification, thus that $\mathcal{C}(x) - \mathcal{C}(y) = \mathcal{C}(x-y)$:

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

This expression is nearly the same as in Equation (S15), 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 $\mathbb{E}\left[\left\|h_{k+1}^{i,a} - h_{k+1}^{i,b}\right\|^2\right]$ using the same line of proof as for Equation (S16), resulting in:

$$
\begin{aligned}
\frac{1}{pN^2} \sum_{i=0}^N \mathbb{E}\left[\left\|h_{k+1}^{a,i} - h_{k+1}^{b,i}\right\|^2 \;\middle|\; \mathcal{H}_k\right] &\leq \left(1 + p\left(2\alpha^2\omega_{\mathcal{C}}^{\mathrm{up}} + 2\alpha^2 - 3\alpha\right)\right) \frac{1}{p^2N^2} \sum_{i=0}^N \mathbb{E}\left[\left\|h_k^{a,i} - h_k^{b,i}\right\|^2 \;\middle|\; \mathcal{H}_k\right] \\
&+ 2(2\alpha^2\omega_{\mathcal{C}}^{\mathrm{up}} + 2\alpha^2 - \alpha)\frac{1}{pN^2} \sum_{i=0}^N \mathbb{E}\left[\left\|g_{k+1}^i(w_k^a) - g_{k+1}^i(w_k^b)\right\|^2 \;\middle|\; \mathcal{H}_k\right].
\end{aligned}
$$

Combining both equations, and using Assumptions 1 and 2 and Equation (S28) we get, under conditions on the learning rates α, γ similar to the ones in Theorems S4 and S5, that

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

And by induction:

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

□

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 Dieuleveut et al. [2019]. This concludes the proof of Proposition S9.

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 7 (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:*

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

Assumption 8 (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 9 (Lower bound on the compression operator's variance)**.** *There exists a constant $\omega_{\mathcal{C}} \in \mathbb{R}^*$ such that the compression operators $\mathcal{C}_{\mathrm{up}}$ and $\mathcal{C}_{\mathrm{dwn}}$ verify the following property:*

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

This last assumption is valid for Stochastic Sparsification.

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 10 (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 $\mathbb{E}[\|w_k - w_*\|^2]$ 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 and w following distributions $\delta_{w_0} R_{\gamma,v}^k$ and $\pi_{\gamma,v}$, and coupled such that they achieve the equality in Equation (S26):

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

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 (S26) at line (ii), and finally Proposition S9 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 $\Xi_{k+1}(w_k,h_k)$ the *global* noise, defined by

$$\Xi_{k+1}(w_k,h_k)=\nabla F(w_k)-\mathcal{C}_{\mathrm{dwn}}\left(\frac{1}{pN}\sum_{i\in S_k}\mathcal{C}_{\mathrm{up}}(g^i_{k+1}(w_k)-h^i_k)+h^i_k\right),$$

such that $w_{k+1}=w_k-\gamma\nabla F(w_k)+\gamma\Xi_{k+1}(w_k,h_k)$.

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 $\mathrm{Cov}(\Xi(w,h))=\mathbb{E}\left[(\Xi(w,h))^{\otimes 2}\right]$, 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 11. *We assume that:*

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

Remark: with the *Stochastic Sparsification* operator, this assumption can directly be translated into an assumption on the moments and regularity of g^i_k. Note that Point 2 in Assumption 11 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 S17. *Under Assumptions 1 to 5, 10 and 11, we have that*

$$\mathbb{E}_{\pi_{\gamma,v}}\left[\|w-w_*\|^2\right]\underset{\gamma\to 0}{=}\gamma\mathrm{Tr}(A\ \mathrm{Cov}(\Xi(w_*,h_*)))+O(\gamma^2), \tag{S29}$$

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 10), 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:

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

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 11) to get Lemma S17.

More formally, we can rely on Theorem 4 in Dieuleveut et al. [2019]: under Assumptions 1 to 5, 10 and 11, 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 S18. *Under Assumptions 7 to 9, we have that, for any variant v of the algorithm, with the constant E given in Theorem 1 depending on the variant:*

$$\mathrm{Tr}\left(\mathrm{Cov}(\Xi(w_*, h_*))\right) = \Omega\left(\frac{\gamma E}{\mu p N}\right). \tag{S30}$$

Combining Lemmas S17 and S18 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 S10. *Under Assumptions 1 to 5 and 7 to 11, we have that*

$$\mathbb{E}[\|w_k - w_*\|^2] \underset{k\to\infty}{\to} \mathbb{E}_{\pi_{\gamma,v}}\left[\|w - w_*\|^2\right] \underset{\gamma\to 0}{=} \Omega\left(\frac{\gamma E}{\mu p N}\right) + O(\gamma^2), \tag{S31}$$

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 $\mathrm{Tr}(A\mathrm{Cov}(\Xi(w_*, h_*)))$. 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 $(w_k)_k$ and not necessarily the couple $(w_k, (h_k^i)_i)_k$.

We now give the proof of Lemma S18.

Proof. With memory, we have the following:

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

At line (i) we use Assumption 9 for the downlink compression operator with constant $\omega_{\mathcal{C}}^{\mathrm{dwn}}$. At line (ii) we use the fact that $\sum_{i=1}^N h_*^i = \nabla F(w_*) = 0$, the independence of the random variables $\mathcal{C}_{\mathrm{up}}(g_1^i(w_*) - h_*^i), \mathcal{C}_{\mathrm{up}}(g_1^j(w_*) - h_*^j)$ for $i \neq j$ and the fact that they have 0 mean. Line (iii) makes appear the bernouilli variable B_k^i that mark if a worker is activate or not at round k. We use Assumption 9 for the uplink compression operator with constant $\omega_{\mathcal{C}}^{\mathrm{up}}$ in line (iv); and finally Assumption 7 at line (v) to lower bound the variance of the gradients at the optimum. This proof applies to both simple and double compression with $\omega_{\mathcal{C}}^{\mathrm{dwn}} = 0$ or not.

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

Without memory, we have the following computation:

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

At line (i) we use Assumption 9 for the downlink compression operator with constant $\omega_{\mathcal{C}}^{\mathrm{dwn}}$ and the fact that $\sum_{i=1}^N h_*^i = \nabla F(w_*) = 0$, then at line (ii) the independence of the random variables $\mathcal{C}_{\mathrm{up}}(g_1^i(w_*)) - h_*^i$ with mean 0, then a Bias Variance decomposition at line (iii).

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

Next we use Assumption 9 for the uplink compression operator with constant $\omega_{\mathcal{C}}^{\mathrm{up}}$ at line (iv). Line (v) is another Bias-Variance decomposition and we finally conclude by using Assumptions 7 and 8 at line (vi) and reorganizing terms.

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