Preserved central model for faster bidirectional compression in distributed settings

Constantin Philippenko, Aymeric Dieuleveut

Preserved central model for faster bidirectional compression in distributed settings

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

Abstract

We develop a new approach to tackle communication constraints in a distributed learning problem with a central server. We propose and analyze a new algorithm that performs bidirectional compression and achieves the same convergence rate as algorithms using only uplink (from the local workers to the central server) compression. To obtain this improvement, we design MCM, an algorithm such that the downlink compression *only impacts local models*, while the global model is preserved. As a result, and contrary to previous works, the gradients on local servers are computed on *perturbed models*. Consequently, convergence proofs are more challenging and require a precise control of this perturbation. To ensure it, MCM additionally combines model compression with a memory mechanism. This analysis opens new doors, e.g. incorporating worker dependent randomized-models and partial participation.

1 Introduction

Large scale distributed machine learning is widely used in many modern applications [1, 8, 40]. The training is distributed over a potentially large number N of workers that communicate either with a central server [see 23, 33, on federated learning], or using peer-to-peer communication [11, 46, 44].

In this work, we consider a setting using a central server that aggregates updates from remote nodes. Formally, we have 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 distributed convex optimization problem using stochastic gradient algorithms [37, 7]: $\min_{w \in \mathbb{R}^d} F(w)$ with $F(w) = \frac{1}{N} \sum_{i=1}^{N} F_i(w)$, where $(F_i)_{i=1}^N$ is a *local* risk function (empirical risk or expected risk in a streaming framework). This applies to both instances of *distributed* and *federated* learning.

An important issue of those frameworks is the high communication cost between the workers and the central server [21, Sec. 3.5]. This cost is a concern from several points of view. First, exchanging information can be the bottleneck in terms of speed. Second, the data consumption and the bandwidth usage of training large distributed models can be problematic; and furthermore, the energetic and environmental impact of those exchanges is a growing concern. Over the last few years, new algorithms were introduced, compressing messages in the *upload communications* (i.e., from remote devices to the central server) in order to reduce the size of those exchanges [41, 3, 49, 2, 47, 43, 42, 34, 28]. More recently, a new trend has emerged to also compress the *downlink communication*: this is *bidirectional compression*.

The necessity for bidirectional compression can depend on the situation. For example, a single uplink compression could be sufficient in *asymmetric* regimes in which broadcasting a message to N workers ("one to N") is faster than aggregating the information coming from each node ("N to one"). However, in other regimes, e.g. with few machines, where the bottleneck is the transfer time of a heavy model (up to several GB in modern Deep Learning architectures) the downlink communication cannot be disregarded, as the upload and download speed are of the same order [36].

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Furthermore, in a situation in which participants have to systematically download an update (e.g., on their smartphones) to participate in the training, participants would prefer to receive a small size update (compressed) rather than a heavier one. To encompass all situations, we consider algorithms for which the information exchanged is compressed in both directions.

To perform downlink communication, existing bidirectional algorithms [45, 52, 38, 29, 36, 17, 51, 14] first aggregate all the information they have received, compress them and then carry out the broadcast. Both the main "global" model and the "local" ones perform the *same* update with this compressed information. Consequently, the model hold on the central server and the one used on the local workers (to query the gradient oracle) are identical. However, this means that the model on the central server has been artificially *degraded*: instead of using all the information it has received, it is updated with the compressed information.

Here, we focus on *preserving* (instead of *degrading*) the central model: the update made on its side does not depend on the downlink compression. This implies that the local models are *different* from the central model. The local gradients are thus measured on a *"perturbed model"* (or *"perturbed iterate"*): such an approach requires a more involved analysis and the algorithm must be carefully designed to control the deviation between the local and global models [31]. For example, algorithms directly compressing the model or the update would simply not converge.

We propose `MCM` - *Model Compression with Memory* - a new algorithm that 1) preserves the central model, and 2) uses a memory scheme to reduce the variance of the local model. We prove that the convergence of this method is similar to the one of algorithms using only unidirectional compression.

Potential Impact. Proposing an analysis that handles perturbed iterates is the key to unlock three major challenges of distributed learning run with bidirectionally compressed gradients. First, we show that it is possible to improve the convergence rate by sending *different* ***randomized*** *models* to the different workers, this is `Rand-MCM`. Secondly, this analysis also paves the way to deal with partially participating machines: the adaptation of `Rand-MCM` to this framework is straightforward; while adapting existing algorithms [38] to partial participation is not practical. Thirdly, this framework is also promising in terms of business applications, e.g., in the situation of learning with privacy guarantees and *with a trusted central server*. We detail those three possible extensions in Subsection 4.1.

Broader impact. This work is aligned with a global effort to make the usage of large scale Federated Learning sustainable by minimizing its environmental impact. Though the impact of such algorithms is expected to be positive, at least on environmental concerns, cautiousness is still required, as a rebound effect may be observed [15]: having energetically cheaper and faster algorithms may result in an increase of such applications, annihilating the gain made by algorithmic progress.

Contributions. We make the following contributions:

1. We propose a new algorithm `MCM`, combining a memory process to the "preserved" update. To convey the key steps of the proof, we also introduce an auxiliary hypothetical algorithm, `Ghost`.
2. For those algorithms, we carefully control the variance of the local models w.r.t. the global one. We provide a *contraction equation* involving the control on the local model's variance and show that `MCM` achieves the same rate of convergence as single compression in strongly-convex, convex and non-convex regimes. We give a comparisons of `MCM`'s rates with existing algorithms in Table 2.
3. We propose a variant, `Rand-MCM` incorporating diversity into models shared with the local workers and show that it improves convergence for quadratic functions.

This is the first algorithm for double compression to focus on a **preserved central model**. We underline, both theoretically and in practice, that we get the same asymptotic convergence rate for simple and double compression - which is a major improvement. Our approach is one of the first to allow for worker dependent model, and to naturally adapt to worker dependent compression levels.

The rest of the paper is organized as follows: in Section 2 we present the problem statement and introduce `MCM` and `Rand-MCM`. Theoretical results on these algorithms are successively presented in Sections 3 and 4. Finally, we present experiments supporting the theory in Section 5.

2 Problem statement

We consider the minimization problem described in Section 1. In the convex case, we assume there exists an optimal parameter w_*, and denote $F_* = F(w_*)$. 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 gradient oracle can be

Table 1: Features of the main existing algorithms performing compression. e_k^i (resp. E_k) denotes the use of error-feedback at uplink (resp. downlink). h_k^i (resp. H_k) denotes the use of a memory at uplink (resp. downlink). Note that `Dist-EF-SGD` is identical to `Double-Squeeze` but has been developed simultaneously and independently.

	Compr.	e_k^i	h_k^i	E_k	H_k	Rand.	update point
`Qsgd` [3]	one-way						
`ECQ-sgd` [49]	one-way	✓					
`Diana` [34]	one-way		✓				
`Dore` [29]	two-way		✓	✓			degraded
`Double-Squeeze` [45], `Dist-EF-SGD` [52]	two-way	✓		✓			degraded
`Artemis` [36]	two-way		✓				degraded
`MCM`	two-way		✓		✓		non-degraded
`Rand-MCM`	two-way		✓		✓	✓	non-degraded

computed on a mini-batch of size b. This function is then evaluated at point w_k. In the classical centralized framework (without compression), for a learning rate γ, SGD corresponds to:

$$w_{k+1} = w_k - \gamma \frac{1}{N} \sum_{i=1}^{N} g_{k+1}^i(w_k) \,. \tag{1}$$

We now describe the framework used for compression.

2.1 Bidirectional compression framework

Bidirectional compression consists in compressing communications in both directions between the central server and remote devices. We use two different compression operators, respectively $\mathcal{C}_{\rm up}$ and $\mathcal{C}_{\rm dwn}$ to compress the message in each direction. Roughly speaking, the update in eq. (1) becomes:

$$w_{k+1} = w_k - \gamma \mathcal{C}_{\rm dwn}\left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\rm up}(g_{k+1}^i(w_k)) \right) .$$

However, this approach has a major drawback. The central server receives and aggregates information $\frac{1}{N}\sum_{i=1}^N \mathcal{C}_{\rm up}(g_{k+1}^i(w_k))$. But in order to be able to broadcast it back, it compresses it, *before* applying the update. We refer to this strategy as the "degraded update" approach. Its major advantage is simplicity, and it was used in all previous papers performing double compression. Yet, it appears to be a waste of valuable information. In this paper, we update the global model w_{k+1} independently of the downlink compression:

$$\left\{ \begin{array}{l} w_{k+1} = w_k - \gamma \frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\rm up}\left(g_{k+1}^i(\hat{w}_k)\right) \,. \\ \hat{w}_{k+1} = C_{\rm dwn}(w_{k+1}) \end{array} \right. \tag{2}$$

However, bluntly compressing w_{k+1} in eq. (2) hinders convergence, thus the second part of the update needs to be refined by adding a memory mechanism. **We now describe both communication stages of the real `MCM`, which is entirely defined by the following uplink and downlink equations.**

Downlink

$$\left\{ \begin{array}{l} \Omega_{k+1} = w_{k+1} - H_k \,, \\ \widehat{w}_{k+1} = H_k + \mathcal{C}_{\rm dwn}(\Omega_{k+1}) \\ H_{k+1} = H_k + \alpha_{\rm dwn} \mathcal{C}_{\rm dwn}(\Omega_{k+1}). \end{array} \right.$$

Uplink

$$\left\{ \begin{array}{l} \forall i \in [\![1, N]\!], \Delta_k^i = g_{k+1}^i(\widehat{w}_k) - h_k^i \\ w_{k+1} = w_k - \frac{\gamma}{N} \sum_{i=1}^N \mathcal{C}_{\rm up}(\Delta_k^i) + h_k^i \\ h_{k+1}^i = h_k^i + \alpha_{\rm up} \mathcal{C}_{\rm up}(\Delta_k^i). \end{array} \right. \tag{3}$$

Downlink Communication. We introduce a *downlink memory term* $(H_k)_k$, which is available on both workers and central server. The difference Ω_{k+1} between the model and this memory is compressed and exchanged, then the local model is reconstructed from this information. The memory is then updated as defined on left part of eq. (3), with a learning rate $\alpha_{\rm dwn}$.

Introducing this memory mechanism is crucial to control the variance of the local model $\widehat{w}_{k+1}$. To the best of our knowledge `MCM` is the first algorithm that uses such a memory mechanism for downlink compression. This mechanism was introduced by Mishchenko et al. [34] for the uplink compression but with the other purpose of mitigating the impact of heterogeneity, while we use it here to avoid divergence of the local model's variance.

Uplink Communication. The motivation to introduce an uplink memory term h_k^i for each device $i \in [\![1, N]\!]$ is different, and better understood. Indeed, for the uplink direction, this mechanism is only necessary (and then crucial) to handle heterogeneous workers [i.e., with different data distributions, see e.g. 36]. Here, the difference Δ_k^i between the stochastic gradient g_{k+1}^i at the local model $\widehat{w}_k$ (as defined in eq. (3)) and the memory term is compressed and exchanged. The memory is then updated as defined on right part of eq. (3) with a rate $\alpha_{\rm dwn}$.

Remark 1 (Rate $\alpha_{\rm dwn}$)**.** *It is necessary to use $\alpha_{\rm dwn} < 1$. Otherwise, the compression noise tends to propagate and is amplified, because of the multiplicative nature of the compression. In Figure 1 we compare* `MCM`*, with 3 other strategies: compressing only the update, compressing $w_k - \widehat{w}_{k-1}$, (i.e., $\alpha_{\rm dwn} = 1$), and compressing the model (i.e., $H_k = 0$), showing that only* `MCM` *converges.*

Remark 2 (Memory vs Error Feedback)**.** *Error feedback is another technique, introduced by Seide et al. [41]. In the context of double compression, it has been shown to improve convergence for a restrictive class of* contracting *compression operators (which are generally biased) by Zheng et al. [52], Tang et al. [45]. However, we note several differences to our approach. (1) For unbiased operators - as considered in* `Dore`*, it did not lead to any theoretical improvement [Remark 2 in Sec. 4.1., 29]. (2) Moreover, only a fraction (namely $(1 + \omega_{\rm dwn})^{-1}$) of the "error" $w_{k+1} - \hat{w}_{k+1}$ can be preserved in the EF term (see line 18 in algo 1 in Liu et al.). It is thus impossible to recover the central preserved model as a function of the degraded model and the EF term. (3) [52] consider a biased operator and the same compression level for uplink and downlink compression. They also rely on stronger assumptions on the gradient (uniformly bounded) and only tackle the homogeneous case.*

In Table 1 we summarize the main algorithms for compression in distributed training. As downlink communication can be more efficient than uplink, we consider distinct operators $\mathcal{C}_{\rm dwn}, \mathcal{C}_{\rm up}$ and allow the corresponding compressions levels to be distinct: those quantities are defined in Assumption 1.

Assumption 1. *There exists constants $\omega_{\rm up}, \omega_{\rm dwn} \in \mathbb{R}_+^*$, such that the compression operators $\mathcal{C}_{\rm up}$ and $\mathcal{C}_{\rm dwn}$ satisfy the two following properties for all w in $\mathbb{R}^d$: $\mathbb{E}[\mathcal{C}_{\rm up/dwn}(w)] = w$, and $\mathbb{E}[\|\mathcal{C}_{\rm up/dwn}(w) - w\|^2] \leq \omega_{\rm up/dwn}\|w\|^2$. The higher is ω, the more aggressive the compression is.*

We only consider unbiased operators, that encompass sparsification, quantization and sketching. References and a discussion on those operators, and possible extensions of our results to biased operators are provided in Appendix A.1.

Remark 3 (Related work on Perturbed iterate analysis)**.** *The theory of perturbed iterate analysis was introduced by Mania et al. [31] to deal with asynchronous SGD. More recently, it was used by Stich and Karimireddy [42], Gorbunov et al. [14] to analyze the convergence of algorithms with uplink compressions, error feedback and asynchrony. Using gradients at randomly perturbed points can also be seen as a form of randomized smoothing [39], a point we discuss in Appendix A.2.*

2.2 The randomization mechanism, `Rand-MCM`

In this subsection, we describe the key feature introduced in `Rand-MCM`: *randomization*. It consists in performing an independent compression for each device instead of performing a single one for all of them. As a consequence, each worker holds a different model centered around the global one. This introduces some supplementary randomness that stabilizes the algorithm. Formally, we will consider N mutually independent compression operators $\mathcal{C}_{{\rm dwn},i}$ instead of a single one $\mathcal{C}_{\rm dwn}$, and the central server will send to the device i at iteration $k+1$ the compression of the difference between its model and the local memory on worker i: $\mathcal{C}_{{\rm dwn},i}(w_{k+1} - H_k^i)$. The tradeoffs associated with this modification are discussed in Section 4.

The pseudocode of `Rand-MCM` is given in Algorithm 1 in Appendix A. It incorporates all components described above: 1) the bidirectional compression, 2) the model update using the non-degraded point, 3) the two memories, 4) the up and down compression operators, 5) the randomization mechanism.

3 Assumptions and Theoretical analysis

We make standard assumptions on $F : \mathbb{R}^d \to \mathbb{R}$. We first assume that the loss function F is smooth.

Assumption 2 (Smoothness)**.** *F is twice continuously differentiable, and is L-smooth, that is for all vectors w_1, w_2 in $\mathbb{R}^d$: $\|\nabla F(w_1) - \nabla F(w_2)\| \leq L\|w_1 - w_2\|$.*

Results in Section 3 are provided in a convex, strongly-convex and non-convex setting.

Assumption 3 (Strong convexity)**.** *F is μ-strongly convex (or convex if $\mu = 0$), that is for all vectors w_1, w_2 in $\mathbb{R}^d$: $F(w_2) \geq F(w_1) + (w_2 - w_1)^T \nabla F(w_1) + \frac{\mu}{2}\|w_2 - w_1\|_2^2$.*

Next, we present the assumption on the stochastic gradients.

Assumption 4 (Noise over stochastic gradients computation)**.** *The noise over stochastic gradients for a mini-batch of size b, is uniformly bounded: there exists a constant $\sigma \in \mathbb{R}_+$, such that for all k in $\mathbb{N}$, for all i in $[\![1, N]\!]$ and for all w in $\mathbb{R}^d$ we have: $E[\|g_k^i(w) - \nabla F(w)\|^2] \leq \sigma^2/b$.*

We here provide guarantees of convergence for `MCM`. `MCM` incorporates an uplink memory term, designed to handle heterogeneous workers. To highlight our main contributions, that concerns the downlink compression, we present the results in the homogeneous setting, that is with $F_i = F_j$ and $\alpha_{\mathrm{up}} = 0$. Similar results (almost identical, up to constant numerical factors) in to the heterogeneous setting are described in Appendix G. Experiments are also performed on heterogeneous workers. We provide here convergence results in the strongly-convex, then convex case.

Notations and settings. For k in $\mathbb{N}$, we denote $\Upsilon_k = \|w_k - H_{k-1}\|^2$, and define $V_k = \mathbb{E}[\|w_k - w_*\|^2] + 32\gamma L \omega_{\mathrm{dwn}}^2 \mathbb{E}[\Upsilon_k]$, which serves as Lyapunov function. V_k is composed of two terms: the first one controls the quadratic distance to the optimal model, and the second controls the variance of the local models $\hat{w}_k$. For both theorems, we choose $\alpha_{\mathrm{dwn}} = (8\omega_{\mathrm{dwn}})^{-1}$. We denote $\Phi(\gamma) := (1 + \omega_{\mathrm{up}})\left(1 + 64\gamma L \omega_{\mathrm{dwn}}^2\right)$.

Limit learning rate: There exists a maximal learning rate to ensure convergence. More specifically, we define $\gamma_{\max} := \min(\gamma_{\max}^{\mathrm{up}}, \gamma_{\max}^{\mathrm{dwn}}, \gamma_{\max}^{\Upsilon})$, where $\gamma_{\max}^{\mathrm{up}} := (2L\,(1 + \omega_{\mathrm{up}}/N))^{-1}$ corresponds to the classical constraint on the learning rate in the unidirectional regime [see 34, 36], $\gamma_{\max}^{\mathrm{dwn}} := (8L\omega_{\mathrm{dwn}})^{-1}$ is a similar constraint coming from the downlink compression, and $\gamma_{\max}^{\Upsilon} := \left(8\sqrt{2}L\omega_{\mathrm{dwn}}\sqrt{8\omega_{\mathrm{dwn}} + \omega_{\mathrm{up}}/N}\right)^{-1}$ is a combined constraint that arises when controlling the variance term Υ.[1] Overall, this constraints are weaker than in the "degraded" framework [29, 36], in which $\gamma_{\max}^{\mathrm{Dore}} \leq \left(8L(1 + \omega_{\mathrm{dwn}})(1 + \omega_{\mathrm{up}}/N)\right)^{-1}$. Especially, in the regime in which $\omega_{\mathrm{up,dwn}} \to \infty$ and $\omega_{\mathrm{dwn}} \simeq \omega_{\mathrm{up}} \simeq: \omega$, the maximal learning rate for `MCM` is $(L\omega^{3/2})^{-1}$, while it is $(L\omega^2)^{-1}$ in [29, 36]. Our $\gamma_{\max}$ is thus larger by a factor $\sqrt{\omega}$, see Table 2. We define $\widetilde{L}$ such that $\gamma_{\max} = (2\widetilde{L})^{-1}$.

Theorem 1 (Convergence of `MCM` in the homogeneous and strongly-convex case)**.** *Under Assumptions 1 to 4 with $\mu > 0$, for k in $\mathbb{N}$, for any sequence $(\gamma_k)_{k\geq 0} \leq \gamma_{\max}$ we have:*

$$V_k \leq (1 - \gamma_k \mu) V_{k-1} - \gamma_k \mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] + \frac{\gamma_k^2 \sigma^2 \Phi(\gamma_k)}{Nb}\,, \tag{4}$$

Consequently, (1) if $\sigma^2 = 0$ (noiseless case), for $\gamma_k \equiv \gamma_{\max}$ we recover a linear convergence rate: $\mathbb{E}[\|w_k - w_*\|^2] \leq (1 - \gamma_{\max}\mu)^k V_0$; (2) if $\sigma^2 > 0$, taking for all K in $\mathbb{N}$, $\gamma_K = 2/(\mu(K+1) + \widetilde{L})$, for the weighted Polyak-Ruppert average $\bar{w}_K = \sum_{k=1}^K \lambda_k w_{k-1} / \sum_{k=1}^K \lambda_k$, with $\lambda_k := (\gamma_{k-1})^{-1}$,

$$\mathbb{E}\left[F(\bar{w}_K) - F(w_*)\right] \leq \frac{\mu + 2\widetilde{L}}{4\mu K^2}\|w_0 - w_*\|^2 + \frac{4\sigma^2(1 + \omega_{\mathrm{up}})}{\mu K N b}\left(1 + \frac{64 L \omega_{\mathrm{dwn}}^2}{\mu K}\ln(\mu K + \widetilde{L})\right). \tag{5}$$

Limit Variance (Equation (4)). For a constant γ, the variance term (i.e., term proportional to σ^2) in Equation (4) is upper bounded by $\frac{\gamma^2\sigma^2}{Nb}(1 + \omega_{\mathrm{up}})(1 + 64\gamma L\omega_{\mathrm{dwn}}^2)$. The impact of the downlink compression is attenuated by a factor γ. As γ decreases, this makes the limit variance similar to the one of `Diana`, i.e., without downlink compression [34, Eq. 16 in Th. 2] and much lower than the variance for previous algorithms using double compression for which the variance scales quadratically with the compression constants as $\gamma^2\sigma^2(1 + \omega_{\mathrm{up}})(1 + \omega_{\mathrm{dwn}})/N$: (1) for `Dore`, see Corollary 1 in Liu et al. [29] (who indicate $(1 - \rho)^{-1} \geq (1 + \omega_{\mathrm{up}}/N)(1 + \omega_{\mathrm{dwn}})$), (2) for `Artemis` see Table 2 and Th. 3 point 2 in [36], (3) for [14], see Theorem I.1. (with $\gamma D_1' \propto \gamma^2\sigma^2(1 + \omega_{\mathrm{up}})(1 + \omega_{\mathrm{dwn}})/N$).

Bound 5 has a quadratic dependence on ω_{dwn}, but the corresponding term is divided by an extra factor K, the number of iterations. For example in experiments, for *w8a* using quantization with $s = 2^0$, we have $\omega_{\mathrm{dwn}} \simeq 17$, and after only 50 epoch with a batch size $b = 12$, we have $K \simeq 2500$. Hence, the term ω^2/K is vanishing through iterations and we asymptotically recover a rate of convergence equivalent to algorithms using unidirectional compression.

[1] The dependency in $\omega^{3/2}$ is similar to the one obtained by Horváth et al. [18] in unidirectional compression in the non-convex case (Theorem 4).

Convergence and complexity: With a decaying sequence of steps, we obtain a convergence rate scaling as $O(K^{-1})$ in Equation (5), without dependency on the $\omega_{\rm dwn}$ in the dominating term, which only appears in faster decaying terms scaling as K^{-2}. The iteration complexity (i.e., number of iterations to achieve ϵ expected error) is thus at first order $O_{\epsilon\to 0}(\frac{\sigma^2(1+\omega_{\rm up})}{\mu\epsilon Nb})$. Again, this matches the complexity of Diana [18, see Theorem 1 and Corollary 1] and is smaller by a factor $1+\omega_{\rm dwn}$ than the one of `Artemis`, `Dore`, `DIANAsr-DQ` (see Corollary I.1. in [14]). Next, we give a convergence result in the convex case.

Theorem 2 (Convergence of `MCM`, convex case)**.** *Under Assumptions 1 to 4 with $\mu = 0$. For all $k > 0$, for any $\gamma \leq \gamma_{\max}$, we have, for $\bar{w}_k = \frac{1}{k}\sum_{i=0}^{k-1} w_i$,*

$$\gamma\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] \leq V_{k-1} - V_k + \frac{\gamma^2\sigma^2\Phi(\gamma)}{Nb} \Longrightarrow \mathbb{E}[F(\bar{w}_k) - F_*] \leq \frac{V_0}{\gamma k} + \frac{\gamma\sigma^2\Phi(\gamma)}{Nb}\,. \quad (6)$$

Consequently, for K in $\mathbb{N}$ large enough, a step-size $\gamma = \sqrt{\frac{\|w_0-w_*\|^2 Nb}{(1+\omega_{\rm up})\sigma^2 K}}$, we have:

$$\mathbb{E}[F(\bar{w}_K) - F_*] \leq 2\sqrt{\frac{\|w_0 - w_*\|^2 (1+\omega_{\rm up})\sigma^2}{NbK}} + O(K^{-1}). \quad (7)$$

Moreover if $\sigma^2 = 0$ (noiseless case), we recover a faster convergence: $\mathbb{E}[F(\bar{w}_K) - F_*] = O(K^{-1})$.

Limit Variance (Eq. (6)). The variance term is identical to the strongly-convex case.

Convergence and complexity (Equation (7)). The downlink compression constant only appears in the second-order term, scaling as $1/K$. In other words, the convergence rate is equivalent to the convergence rate of `Diana`, in the non-strongly-convex. As K increases, this complexity scales as $\frac{(1+\omega_{\rm up})}{n\epsilon^2}$ independently of the downlink compression. Again, for previous algorithms with double compression the complexity is at least $O\left(\frac{(1+\omega_{\rm up})(1+\omega_{\rm dwn})}{n\epsilon^2}\right)$ (see Corollary I.2 in [14]).

Control of the variance of the local model.

We here present the backbone Lemma of `MCM`'s proof. It allows to control the variance of the local model $\mathbb{E}[\|\hat{w}_k - w_k\|^2 | w_k]$ (which is upper-bounded by $\omega_{\rm dwn}\mathbb{E}[\|\Upsilon_k\|^2 | w_k]$) and to build the Lyapunov function defined in Theorems 1 and 2.

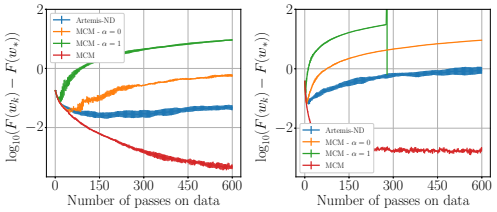


Figure 1: Comparing `MCM` on two datasets with three other algorithms using a non-degraded update, $\gamma = 1/L$. `Artemis-ND` stands for `Artemis` with a non-degraded update.

This result highlights the impact of the downlink memory term. Without memory, i.e., with $\alpha_{\rm dwn} = 0$, the variance of the local model $\|\hat{w}_k - w_k\|^2$ increases with the number of iterations. On the other hand, if $\alpha_{\rm dwn}$ is too large (close to 1), this variance diverges. This behavior is illustrated on two real datasets on Figure 1. This phenomenon is similar to the divergence observed in frameworks involving error feedback, when the compression operator is not contractive.

Theorem 3. *Consider the `MCM` update as in eq. (2). Under Assumptions 1, 2 and 4 with $\mu = 0$, if $\gamma \leq (8\omega_{\rm dwn}L)^{-1}$ and $\alpha \leq (4\omega_{\rm dwn})-1$, then for all k in $\mathbb{N}$:*

$$\mathbb{E}\left[\Upsilon_k\right] \leq \left(1 - \frac{\alpha_{\rm dwn}}{2}\right)\mathbb{E}\left[\Upsilon_{k-1}\right] + 2\gamma^2\left(\frac{1}{\alpha_{\rm dwn}} + \frac{\omega_{\rm up}}{N}\right)\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] + \frac{2\gamma^2\sigma^2(1+\omega_{\rm up})}{Nb}\,.$$

This bound provides a recursive control on Υ_k. Beyond the $(1-\alpha_{\rm dwn})$ contraction, the bound comprises the squared-norm of the gradient at the previous perturbed iterate, and a noise term.

Summary of rates. In Table 2, we summarize the rates and complexities, and maximal learning rate for `Diana`, `Artemis`, `Dore` and `MCM`. For simplicity, we ignore absolute constants, and provide asymptotic values for large $\omega_{\rm up}$, $\omega_{\rm dwn}$, and complexities for $\epsilon \to 0$.

Proof in the heterogeneous case. To extend Theorems 1 to 3 in the heterogeneous setting for a convex objective (Appendix G), we assume that there exists a constant B in $\mathbb{R}_+$, s.t.:

Table 2: Summary of rates on the initial condition, limit variance, asympt. complexities and $\gamma_{\max}$.

Problem		Diana	Artemis, Dore	MCM, Rand-MCM
	$L\gamma_{\max} \propto$ Lim. var. $\propto \gamma^2\sigma^2/n\times$	$1/(1+\omega_{\rm up})$ $(1+\omega_{\rm up})$	$1/(1+\omega_{\rm up})(1+\omega_{\rm dwn})$ $(1+\omega_{\rm up})(1+\omega_{\rm dwn})$	$1/(1+\omega_{\rm dwn})\sqrt{1+\omega_{\rm up}} \wedge 1/(1+\omega_{\rm up})$ $(1+\omega_{\rm up})(1+\gamma L\omega_{\rm dwn}^2)$
Str.-convex	Rate on init. cond. (SC) Complexity	$(1-\gamma\mu)^k$ $(1+\omega_{\rm up})/\mu\epsilon N$	$(1-\gamma\mu)^k$ $(1+\omega_{\rm dwn})(1+\omega_{\rm up})/\mu\epsilon N$	$(1-\gamma\mu)^k$ $(1+\omega_{\rm up})/\mu\epsilon N$
Convex	Complexity	$(\omega_{\rm up}+1)/\epsilon^2$	$(1+\omega_{\rm up})(1+\omega_{\rm dwn})/\epsilon^2$	$(\omega_{\rm up}+1)/\epsilon^2$

$\frac{1}{N}\sum_{i=0}^N \|\nabla F_i(w_*)\|^2 = B^2$. We further define $\Xi_k = \frac{1}{N^2}\sum_{i=1}^N \left\|h_k^i - \nabla F_i(w_*)\right\|^2$, where for all i in $[\![1, N]\!]$. This term is recursively controled [34, 36] and combined into the Lyapunov function.

Proofs. To convey the best understanding of the theorems and the spirit of the proof, we introduce a `Ghost` algorithm (impossible to implement) in Appendix D.1. A sketch of the proof describes the main steps in the case of `Ghost`, those steps are similar for `MCM`. Fundamentally, our proof relies on a tight analysis, related to perturbed iterate analysis [31]. Proofs of Theorems 1 to 3 are given in Appendix E. Th. S11 in Appendix E.4 ensures convergence for a non-convex F. Note that the proof for non-convex follows a different approach than the one in Theorems 1 and 2.

As mentioned in the introduction, our analysis of perturbed iterate in the context of double compression opens new directions: in particular, it opens the door to handling a different model for each worker. In the next section, we detail those possibilities, and provide theoretical guarantees for `Rand-MCM`, the variant of `MCM` in which instead of sending the same model to all workers, the compression noises are mutually *independent*.

Remark 4 (Communication budget)**.** *How to split a given communication budget between uplink and downlink to optimize the convergence is an open question which is intrinsically related to the situation. Indeed it depends on many factors like the selected operators of compression, the upload/downlink speed or the number of participating workers at each iteration. However, our approach provides some insights on this question. Because asymptotically the impact of double compression is marginal, for a fixed budget, Theorem 2 suggests to strongly compress on the downlink direction (which leads to a large ω_{dwn}), but to perform a weaker compression in the uplink direction.*

4 Extension to `Rand-MCM`

4.1 Communication and convergence trade-offs

In `Rand-MCM`, we leverage the fact that the compressions used for each worker need not to be identical. On the contrary, it is possible to consider *independent* compressions. By doing so, we reduce the impact of the downlink compression.

The relevance of such a modification depends on the framework: while the convergence rate will be improved, the computational time can be slightly increased. Indeed, N compressions need to be computed instead of one: however, this computational time is typically not a bottleneck w.r.t. the communication time. A more important aspect is the communication cost. While the size of each message will remain identical, a different message needs to be sent to each worker. That is, we go from a "one to N" configuration to N "one to one" communications. While this is a drawback, it is not an issue when the bandwidth/transfer time are the bottlenecks, as `Rand-MCM` will result in a better convergence with almost no cost. Furthermore, we argue that handling worker dependent models is essential for several major applications. `Rand-MCM` can directly be adapted to those frameworks.

1. Worker dependent compression. A first simple situation is the case in which workers are allowed to choose the size (or equivalently the compression level) of their updates.

2. Partial participation (PP). Similarly, having N different messages to send to each worker may be unavoidable in the case of *partial participation* of the workers. This is a key feature in Federated Learning frameworks [33]. In the classical distributed framework (without downlink constraints) it is easy to deal with it, as each available worker just queries the global model to compute its gradient on it [see for example 17]. On the other hand, for bidirectional compression, to ensure that all the local models match the central model, the adaptation to partial participation relies on a *synchronization step*. During this step, each worker that has not participated in the last S steps receives the last S corresponding messages as long as it costs less to send this sequence than a full uncompressed model. This is described in the description of the adaptation to partial participation in [36], in the remark preceding Eq. (20) in [38] and by Tang et al. [45, v2 on arxiv for the distributed case], who use

a buffer. On the contrary, Rand-MCM naturally handles a different model, memory and update per worker. The adaptation to partial participation is thus straightforward. Though theoretical results are out of the scope of this paper, we provide experiments on PP in Appendix B.1.1 and fig. 4.

One drawback is the necessity to store the N memories $(H_k^i)_{i\in[N]}$ instead of one, which results in an additional memory cost. To circumvent this issue we propose two independent solutions. 1) Keep and use a single memory $\bar{H}_k = N^{-1}\sum_{i=1}^N H_k^i$ (as suggested in [36]). It is then necessary to periodically reset the local memories H_k^i on all workers to the averaged value $\bar{H}_k$ (rarely enough not to impact the communication budget). This is illustrated in fig. 4. 2) Use Rand-MCM with an arbitrary number of groups $G \ll N$ of workers. In each group $\mathcal{G}_g$, $g \in [G]$, all workers share the same memory (H_k^g) and receive the same update $\mathcal{C}_{\mathrm{dwn},g}(w_{k+1} - H_k^g)$. We call this algorithm Rand-MCM-G.

Remark 5 (Protecting the global model from honest-but-curious clients)**.** *Another business advantage of MCM and Rand-MCM is that providing degraded models to the participants can be used to guarantee privacy, or to ensure the workers participate in good faith, and not only to obtain the model. This issue of detecting ill-intentioned clients (free-riders) that want to obtain the model without actually contributing has been studied by Fraboni et al. [13].*

4.2 Theoretical results

In this Section, we provide two main theoretical results for Rand-MCM. First Theorem 4 ensures that the theoretical guarantees are at least as good for Rand-MCM as for MCM. Then, in Theorem 5, we provide convergence result for both MCM and Rand-MCM in the case of quadratic functions.

Theorem 4. *Theorems 1 to 3 are valid for Rand-MCM and Rand-MCM-G.*

The improvement in Rand-MCM comes from the fact that we are ultimately averaging the gradients at several random points, reducing the variance coming from this aspect. The goal is obviously to reduce the impact of ω_{dwn}. Keeping in mind that the dominating term in the rate is independent of ω_{dwn}, *we can thus only expect to reduce the second-order term*. Next, the uplink compression noise increases with the variance of the randomized model, which will not be directly reduced by Rand-MCM. As a consequence, we only expect the improvement to be visible in the part of the second-order term that does not depend on ω_{up} (that is, the effect would be the most significant if ω_{up} is small or 0).

This intuition is corroborated by the following result, in which we show that the convergence is improved when adding the randomization process for a quadratic function. Extending the proof beyond quadratic functions is possible, though it requires an assumption on third or higher order derivatives of F (e.g., using self-concordance [5]) to control of $\mathbb{E}\left[||\nabla F(\widehat{w}_{k-1}) - \mathbb{E}[\nabla F(\widehat{w}_{k-1})]||^2 \,\middle|\, w_{k-1}\right]$.

Theorem 5 (Convergence in the quadratic case)**.** *Under Assumptions 1 to 4 with $\mu = 0$, if the function is quadratic, after running $K > 0$ iterations, for any $\gamma \leq \gamma_{\max}$, and we have*

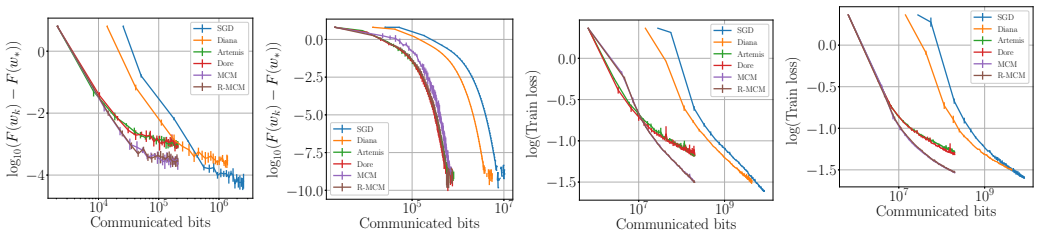
$$\mathbb{E}[F(\bar{w}_K) - F_*] \leq \frac{V_0}{\gamma K} + \frac{\gamma\sigma^2\Phi^{\mathrm{Rd}}(\gamma)}{Nb}\,,$$

with $\Phi^{\mathrm{Rd}}(\gamma) = (1+\omega_{\mathrm{up}})\left(1 + \frac{4\gamma^2L^2\omega_{\mathrm{dwn}}}{K}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\mathrm{up}}}{N}\right)\right)$ *and* $\mathbf{C} = N$ *for Rand-MCM,* $\mathbf{C} = G$ *Rand-MCM-G, and* $\mathbf{C} = 1$ *for MCM.*

This result is derived in Appendix F. We can make the following comments: (1) The convergence rate for quadratic functions is slightly better than for smooth functions. More specifically, the right hand term in Φ is multiplied by an additional $\gamma\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\mathrm{up}}}{N}\right)$ (w.r.t. Theorem 2), which is decaying at the same rate as γ. Besides, the proof for Rand-MCM is substantially modified, as $\mathbb{E}[\nabla F(\widehat{w}_{k-1})]$ is an unbiased estimator of $\nabla F(w_{k-1})$. (2) Moreover, the randomization in Rand-MCM (resp. Rand-MCM-G) further reduces by a factor N (resp. G) this term. Depending on the relative sizes of ω_{up} and N, this can lead to a significant improvement up to a factor of N. In practice the impact of Rand-MCM is noticeable, as illustrated in the following experiments.

5 Experiments

In this section, we illustrate the validity of the theoretical results given in the previous section on both synthetic and real datasets, on (1) least-squares linear regression (LSR), (2) logistic regression (LR), and (3) non-convex deep learning. We compare MCM with classical algorithms used in distributed settings: Diana, Artemis, Dore and of course the simplest setting - SGD, which is the baseline.



(a) $\sigma^2 \neq 0, \gamma = (L\sqrt{k})^{-1}$ (b) $\sigma^2 = 0, \gamma = L^{-1}$ (c) MNIST with a CNN (d) FE-MNIST with a CNN

Figure 2: Convergence on neural networks.

In these experiments, we provide results on the log of the excess loss $F(w_k) - F_*$, averaged on 5 runs (resp. 2) in convex settings (resp. deep learning), with errors bars displayed on each figure (but not in the "zoom square"), corresponding to the standard deviation of $\log_{10}(F(w_k) - F_*)$. On Figure 3, the X-axis is respectively the number of iterations and the number of bits exchanged.

Each experiment has been run with $N = 20$ workers using stochastic scalar quantization [3], w.r.t. 2-norm. To maximize compression, we always quantize on a single level ($s = 2^0$), unless for PP ($s = 2^1$) and neural network (the value of s depends on the dataset).

We used 9 different datasets.

- One toy dataset devoted to linear regression in an homogeneous setting. This toy dataset allows to illustrate `MCM` properties in a simple framework, and in particular to ilustrate that when $\sigma^2 = 0$, we recover a linear convergence[2], see Figure 2b.
- Five datasets commonly used in convex optimization (a9a, quantum, phishing, superconduct and w8a); see Table S1 for more details. Experiments were conducted with heterogeneous workers obtained by clustering (using *TSNE* [30]) the input points.
- Four dataset in a non-convex settings (CIFAR10, Fashion-MNIST, FE-MNIST, MNIST); see Table S2 for more details.

All experiments are performed without any tuning of the algorithms, (e.g., with the same learning rate for all algorithms and without reducing it after a certain number of epochs). Indeed, our goal is to show that our method achieves a performance close to the unidirectional-compression framework (`Diana`), while performing an important downlink compression. More details about experiments can be found in Appendix B.

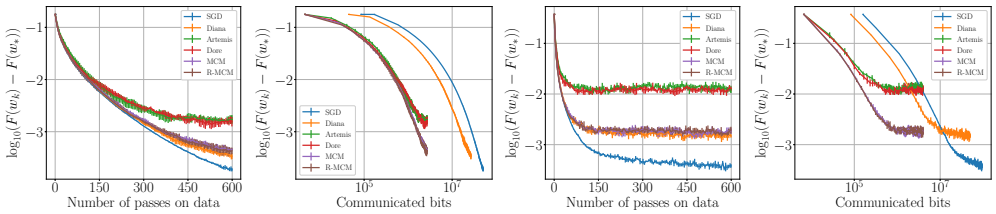
On Figure 3, we display the excess loss for quantum and a9a w.r.t. the number of iteration and number of communicated bits. The plots of phising, superconduct and w8a are not provided but can be found on our github repository. We only report their excess loss after 450 iterations in Table 3.

Table 3: `MCM`- convex experiments, b is the batch size

Excess loss after 450 epochs	`SGD`	`Diana`	`MCM`	`Dore`	Ref
a9a ($b = 50$)	-3.5	-2.7	-2.7	-1.8	[10]
quantum ($b = 400$)	-3.4	-3.2	-3.2	-2.6	[9]
phishing ($b = 50$)	-3.7	-3.5	-3.4	-2.7	[10]
superconduct ($b = 50$)	-1.6	-1.6	-1.55	-1.45	[16]
w8a ($b = 12$)	-3.5	-3.0	-2.5	-1.75	[10]
Compression	no	uni-dir	bi-dir	bi-dir	

Saturation level. All experiments are performed with a *constant learning rate* γ to observe the bias (initial reduction) and the variance (saturation level) independently. Stochastic gradient descent results in a fast convergence during the first iterations, and then reaches a saturation at a given level proportional to σ^2. Theorem 2 states that the variance of `MCM` is proportional to ω_{up}, this is experimentally observed on Tables 3 and 4 and figs. 2 and 3: `MCM` meets `Diana` while `Artemis` and `Dore` saturate at a higher level (scaling as $\omega_{\text{up}} \times \omega_{\text{dwn}}$). These trade-offs are preserved with optimized learning rates.

[2]Even stronger, we show in experiments that we recover a linear rate if we have $\sigma_* = 0$ (the noise over stochastic gradient computation at the optimum point $w_\star$).



(a) Quantum in #iter. (b) Quantum in #bits (c) A9A in #iter. (d) A9A in #bits

Figure 3: Experiments on real dataset with $\gamma = 1/L$, quantization with $s = 1$, LSR (a,b), LR (c,d).

Linear convergence when $\sigma^2 = 0$. The six algorithms present a linear convergence when $\sigma^2 = 0$. This is illustrated by Figure 2b: we ran experiments with a full gradient descent. Note that in these settings `MCM` has a slightly worse performance than other methods; however, this slow-down is compensated by `Rand-MCM`.

Impact of randomization. The impact of randomization is noticeable on Figures 2b and S5b. Randomization helps to stabilise convergence of it reduces the variance of the runs and when $\sigma^2 = 0$, it performs identically to `SGD`. Figure 4 illustrates the impact of using a single memory, instead of N, to alleviate the memory cost in the PP setting (Subsection 4.1), with or without periodic reset. Without reset, performance are slightly degraded, but with it, we recover previous results.

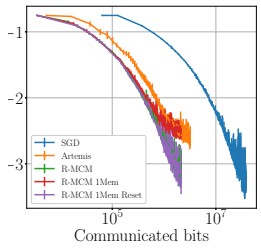


Figure 4: `Rand-MCM` (PP) on quantum with a *single memory* ($s = 2$).

Deep learning. Table 4 and figs. 2c and 2d illustrate experiments with neural networks, details on dataset settings and networks architecture are given in Appendix B.2. Again, `MCM` meets `Diana` rates as stated by Theorem S11 (theorem in the non-convex case).

Table 4: Accuracy and train loss in non-convex experiments, detailed settings can be found in Table S2.

	Algorithm	MNIST	Fashion MNIST	FE-MNIST	CIFAR-10
Accuracy after 300 epochs	`SGD`:	99.0%	92.4%	99.0%	69.1%
	`Diana`:	98.9%	92.4%	98.9%	64.0%
	`MCM`:	98.8%	90.6%	98.9%	63.5%
	`Artemis`:	97.9%	86.7%	98.3%	54.8%
	`Dore`:	97.9%	87.9%	98.5%	56.3%
Train loss after 300 epochs	`SGD`:	0.025	0.093	0.026	0.909
	`Diana`:	0.034	0.141	0.031	1.047
	`MCM`:	0.033	0.209	0.030	1.096
	`Artemis`:	0.075	0.332	0.052	1.342
	`Dore`:	0.072	0.300	0.048	1.292

Overall, these experiments show the benefits of `MCM` and `Rand-MCM`, that reach the saturation level of `Diana` while exchanging at 10x to 100x fewer bits. More experiments with partial participation for `Rand-MCM` are given in Appendix B.1.1. All the code is provided on our github repository.

6 Conclusion

In this work, we propose a new algorithm to perform bidirectional compression while achieving the convergence rate of algorithms using compression in a single direction. One of the main application of this framework is Federated Learning. With `MCM` we stress the importance of not degrading the global model. In addition, we add the concept of randomization which allows to reduce the variance associated with the downlink compression. The analysis of `MCM` is challenging as the algorithm involves perturbed iterates. Proposing such an analysis is the key to unlocking numerous challenges in distributed learning, e.g., proposing practical algorithms for partial participation, incorporating privacy-preserving schemes *after* the global update is performed, dealing with local steps, etc. This approach could also be pivotal in non-smooth frameworks, as it can be considered as a weak form of randomized smoothing.

Acknowledgments

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

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: a system for large-scale machine learning. In *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*, OSDI'16, pages 265–283, USA, November 2016. USENIX Association. ISBN 978-1-931971-33-1.

[2] Naman Agarwal, Ananda Theertha Suresh, Felix Xinnan X Yu, Sanjiv Kumar, and Brendan McMahan. cpSGD: Communication-efficient and differentially-private distributed SGD. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7564–7575. Curran Associates, Inc., 2018.

[3] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding. *Advances in Neural Information Processing Systems*, 30:1709–1720, 2017.

[4] Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cedric Renggli. The Convergence of Sparsified Gradient Methods. *Advances in Neural Information Processing Systems*, 31:5973–5983, 2018.

[5] Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4(none):384–414, January 2010. ISSN 1935-7524, 1935-7524. doi: 10.1214/09-EJS521. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

[6] Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On Biased Compression for Distributed Learning. *arXiv:2002.12410 [cs, math, stat]*, February 2020. arXiv: 2002.12410.

[7] Léon Bottou. Large-Scale Machine Learning with Stochastic Gradient Descent. In Yves Lechevallier and Gilbert Saporta, editors, *Proceedings of COMPSTAT'2010*, pages 177–186, Heidelberg, 2010. Physica-Verlag HD. ISBN 978-3-7908-2604-3. doi: 10.1007/978-3-7908-2604-3_16.

[8] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A Benchmark for Federated Settings. *arXiv:1812.01097 [cs, stat]*, December 2019. arXiv: 1812.01097.

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

[10] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, May 2011. ISSN 2157-6904. doi: 10.1145/1961189.1961199.

[11] Igor Colin, Aurelien Bellet, Joseph Salmon, and Stéphan Clémençon. Gossip Dual Averaging for Decentralized Optimization of Pairwise Functions. In *International Conference on Machine Learning*, pages 1388–1396. PMLR, June 2016. ISSN: 1938-7228.

[12] John C. Duchi, Peter L. Bartlett, and Martin J. Wainwright. Randomized Smoothing for Stochastic Optimization. *SIAM Journal on Optimization*, 22(2):674–701, January 2012. ISSN 1052-6234. doi: 10.1137/110831659. Publisher: Society for Industrial and Applied Mathematics.

[13] Yann Fraboni, Richard Vidal, and Marco Lorenzi. Free-rider attacks on model aggregation in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1846–1854. PMLR, 2021.

[14] Eduard Gorbunov, Dmitry Kovalev, Dmitry Makarenko, and Peter Richtarik. Linearly Converging Error Compensated SGD. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20889–20900. Curran Associates, Inc., 2020.

[15] M. J. Grubb. Communication Energy efficiency and economic fallacies. *Energy Policy*, 18(8): 783–785, October 1990. ISSN 0301-4215. doi: 10.1016/0301-4215(90)90031-X.

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

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

[18] Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic Distributed Learning with Gradient Quantization and Variance Reduction. *arXiv:1904.05115 [math]*, April 2019. arXiv: 1904.05115.

[19] Rui Hu, Yanmin Gong, and Yuanxiong Guo. Sparsified Privacy-Masking for Communication-Efficient and Privacy-Preserving Federated Learning. *arXiv:2008.01558 [cs, stat]*, August 2020. arXiv: 2008.01558.

[20] Nikita Ivkin, Daniel Rothchild, Enayat Ullah, Vladimir Braverman, Ion Stoica, and Raman Arora. Communication-efficient Distributed SGD with Sketching. *Advances in Neural Information Processing Systems*, 32:13144–13154, 2019.

[21] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and Open Problems in Federated Learning. *arXiv:1912.04977 [cs, stat]*, December 2019. arXiv: 1912.04977.

[22] Sarit Khirirat, Sindri Magnússon, Arda Aytekin, and Mikael Johansson. Communication Efficient Sparsification for Large Scale Machine Learning. *arXiv:2003.06377 [math, stat]*, March 2020. arXiv: 2003.06377.

[23] Jakub Konečný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. *arXiv:1610.02527 [cs]*, October 2016. arXiv: 1610.02527.

[24] Alex Krizhevsky, Geoffrey Hinton, and others. Learning multiple layers of features from tiny images. 2009. Publisher: Citeseer.

[25] Loïc Lannelongue, Jason Grealey, and Michael Inouye. Green algorithms: Quantifying the carbon footprint of computation. *Advanced Science*, page 2100707, 2021. Publisher: Wiley Online Library.

[26] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. ISSN 1558-2256. doi: 10.1109/5.726791. Conference Name: Proceedings of the IEEE.

[27] Tian Li, Zaoxing Liu, Vyas Sekar, and Virginia Smith. Privacy for Free: Communication-Efficient Learning with Differential Privacy Using Sketches. *arXiv:1911.00972 [cs, stat]*, December 2019. arXiv: 1911.00972 version: 2.

[28] Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtarik. Acceleration for Compressed Gradient Descent in Distributed and Federated Optimization. In *International Conference on Machine Learning*, pages 5895–5904. PMLR, November 2020. ISSN: 2640-3498.

[29] Xiaorui Liu, Yao Li, Jiliang Tang, and Ming Yan. A Double Residual Compression Algorithm for Efficient Distributed Learning. In *International Conference on Artificial Intelligence and Statistics*, pages 133–143, June 2020. ISSN: 1938-7228 Section: Machine Learning.

[30] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. ISSN ISSN 1533-7928.

[31] Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I. Jordan. Perturbed Iterate Analysis for Asynchronous Stochastic Optimization. *arXiv:1507.06970 [cs, math, stat]*, March 2016. arXiv: 1507.06970.

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

[33] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, April 2017. ISSN: 2640-3498.

[34] Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed Learning with Compressed Gradient Differences. *arXiv:1901.09269 [cs, math, stat]*, June 2019. arXiv: 1901.09269.

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

[36] Constantin Philippenko and Aymeric Dieuleveut. Artemis: tight convergence guarantees for bidirectional compression in Federated Learning. *arXiv:2006.14591 [cs, stat]*, November 2020. arXiv: 2006.14591.

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

[38] Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and Communication-Efficient Federated Learning From Non-i.i.d. Data. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2019. ISSN 2162-2388. doi: 10.1109/TNNLS.2019.2944481. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

[39] Kevin Scaman, Francis Bach, Sebastien Bubeck, Laurent Massoulié, and Yin Tat Lee. Optimal Algorithms for Non-Smooth Distributed Optimization in Networks. *Advances in Neural Information Processing Systems*, 31:2740–2749, 2018.

[40] Frank Seide and Amit Agarwal. CNTK: Microsoft's Open-Source Deep-Learning Toolkit. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 2135, New York, NY, USA, August 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2945397.

[41] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Fifteenth Annual Conference of the International Speech Communication Association*. Citeseer, 2014.

[42] Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for SGD with delayed gradients and compressed updates. *Journal of Machine Learning Research*, 21:1–36, 2020.

[43] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with Memory. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 4447–4458. Curran Associates, Inc., 2018.

[44] Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. D^2: Decentralized Training over Decentralized Data. In *International Conference on Machine Learning*, pages 4848–4856. PMLR, July 2018. ISSN: 2640-3498.

[45] Hanlin Tang, Chen Yu, Xiangru Lian, Tong Zhang, and Ji Liu. DoubleSqueeze: Parallel Stochastic Gradient Descent with Double-pass Error-Compensated Compression. In *International Conference on Machine Learning*, pages 6155–6165. PMLR, May 2019. ISSN: 2640-3498.

[46] Paul Vanhaesebrouck, Aurélien Bellet, and Marc Tommasi. Decentralized Collaborative Learning of Personalized Models over Networks. In *Artificial Intelligence and Statistics*, pages 509–517. PMLR, April 2017. ISSN: 2640-3498.

[47] Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient Sparsification for Communication-Efficient Distributed Optimization. *Advances in Neural Information Processing Systems*, 31:1299–1309, 2018.

[48] Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai 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.

[49] Jiaxiang Wu, Weidong Huang, Junzhou Huang, and Tong Zhang. Error Compensated Quantized SGD and its Applications to Large-scale Distributed Optimization. In *International Conference on Machine Learning*, pages 5325–5333. PMLR, July 2018. ISSN: 2640-3498.

[50] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv:1708.07747 [cs, stat]*, September 2017. arXiv: 1708.07747.

[51] An Xu, Zhouyuan Huo, and Heng Huang. Training Faster with Compressed Gradient. *arXiv:2008.05823 [cs, stat]*, August 2020. arXiv: 2008.05823.

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

[53] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients. *arXiv:1606.06160 [cs]*, February 2018. arXiv: 1606.06160.

Supplementary material

In this appendix, we provide additional details about our work. First, in Appendix A we give complementary references on operators of compression and on perturbed iterate analysis. We also give the pseudo-code of `Rand-MCM`. Secondly, in Appendix B we enlarge figures provided in Section 5 and complete them with experiments on partial participation and with a comparison between `MCM` and other algorithms using non-degraded updates. The next sections are all devoted to theoretical results. In Appendix C we detail some technical results required to demonstrate Theorems 1 to 5, in Appendix D we highlight the key stages of the demonstration in the easier case of `Ghost`, in Appendix E we completely prove the given guarantees of convergence in three regimes: convex, strongly-convex and non-convex. In Appendix F we show the benefit of `Rand-MCM` compared to `MCM` in the context of quadratic functions. In Appendix G we adapt the proof to the heterogeneous scenario. And finally, in Appendix H we answer to the Neurips checklist.

Contents

A Complementary discussions and references

We give the pseudo-code of `Rand-MCM` in Algorithm 1. It summarizes the algorithm's description given in Section 1.

A.1 Compression Operators

In this section, we give additional details on compression operators (see Assumption 1).

Operators of compression can be biased or unbiased and they may have drastically different impacts on convergence. For instance, if the operator is not contracting, algorithms with error-feedback may diverge. Horváth and Richtárik [17] propose a method to unbiase a biased operator and a general study of biased operator has been carried out by Beznosikov et al. [6]. But in this work, as stated by Assumption 1, we consider only unbiased operators: for instance `s-quantization`.

The choice of the operator of compression is crucial when compressing data. Operators of compression may be classified into three mains categories: 1) sparsification [43, 19, 22, 4, 22, 32] 2) quantization [41, 53, 3, 18, 48] and 3) sketching [20, 27].

Possible Extensions Our analysis could be extended to biased uplink operators, following similar lines of proof as [6].

The extension for the downlink operator seems more difficult as our analysis relies on numerous occurrences on the fact that the expectation of $\widehat{w}_{k-1}$ knowing w_{k-1} is w_{k-1}.

Algorithm 1 Pseudocode of `Rand-MCM`

Input: Mini-batch size b, learning rates $\alpha_{\rm up}, \alpha_{\rm dwn}, \gamma > 0$, initial model $w_0 \in \mathbb{R}^d$ (on all devices), operators $\mathcal{C}_{\rm up}$ and $\mathcal{C}_{\rm dwn}$, $S = [\![1, N]\!]$ the set of devices.
Init.: Memories: $\forall i \in S$, $h_0^i = g_1^i(w_0)$ and $H_{-1}^i = w_0$
Output: Model w_K
for $k = 1, 2, \ldots, K$ **do**
  for each device $i = 1, 2, 3, \ldots, N$ **do**
    Receive $\widehat{\Omega}_{k-1}^i$, and set: $w_{k-1}^i = \widehat{\Omega}_{k-1}^i + H_{k-2}^i$
    Compute $g_k^i(w_{k-1}^i)$ (with mini-batch)
    Update down memory: $H_{k-1}^i = H_{k-2}^i + \alpha_{\rm dwn}\widehat{\Omega}_{k-1}^i$
    Up compr.: $\widehat{\Delta}_{k-1}^i = \mathcal{C}_{\rm up}(g_k^i(w_{k-1}^i) - h_{k-1}^i)$
    Update uplink memory: $h_k^i = h_{k-1}^i + \alpha_{\rm up}\widehat{\Delta}_{k-1}^i$
    Send $\widehat{\Delta}_{k-1}^i$ to central server
  end for
  Receive $(\widehat{\Delta}_{k-1}^i)_{i=1}^N$ from all remote servers
  Compute $\widehat{g_k} = \frac{1}{N}\sum_{i=1}^N \widehat{\Delta}_{k-1}^i + h_{k-1}^i$
  Update up memory: $\forall i \in S, h_k^i = h_{k-1}^i + \alpha_{\rm up}\widehat{\Delta}_{k-1}^i$
  Non-degraded update: $w_k = w_{k-1} - \gamma\widehat{g_k}$
  Down compr.: $\forall i \in S$, $\widehat{\Omega}_k^i = \mathcal{C}_{{\rm dwn},i}(w_k - H_{k-1}^i)$
  Update downlink memory: $H_k^i = H_{k-1}^i + \alpha_{\rm dwn}\widehat{\Omega}_k^i$
  Send $(\widehat{\Omega}_k^i)_{i=1}^N$ to all remote servers
end for

A.2 Relation to Randomized Smoothing

Our approach can also be related to randomized smoothing. Formally, $\nabla F(\widehat{w}_{k-1})$ can be considered as an unbiased gradient of the smoothed function F_ρ at point w_{k-1}, with $F_\rho : w \mapsto \mathbb{E}[F(w + \widehat{w}_{k-1} - w_{k-1})]$. Then $\mathbb{E}\left\langle \nabla F(\widehat{w}_{k-1}),\ w_{k-1} - w_* \right\rangle = \mathbb{E}\left\langle \nabla F_\rho(w_{k-1}),\ w_{k-1} - w_* \right\rangle$. One key aspect is that the condition number μ_ρ/L_ρ of F_ρ is always larger (better) than the one for F. However, the minimum of F_ρ is different and moving, thus the proof techniques from Randomized smoothing are not adapted to a varying noise which distribution is unknown. Providing a theoretical result that quantifies the smoothing impact of `MCM` is an interesting open direction.

Randomized smoothing has been applied to non-smooth problems by Duchi et al. [12]. The aim is to transform a non-smooth function into a smooth function, before computing the gradient. This is achieved by adding a Gaussian noise to the point where the gradient is computed. This mechanism has been applied by Scaman et al. [39] to convex problems. We consider in this work a randomized version of compression: at iteration k in $\mathbb{N}$ each worker i in $[\![1, N]\!]$ receives a noisy estimate $\widehat{w}_k^i$ of the global model w_k kept on central server. Thus, we compute the local gradient at a perturbed point $w_k + \delta_k^i$. Unlike the randomization process as defined by Duchi et al. [12], the noise here is not chosen to improve the function's regularity but results from the compression.

B Experiments

In this section we provide additional details about our experiments. We first give the settings of our experiments in Tables S1 and S2. Next, we describe the numerical results obtained on our 9 datasets. Thirdly, we add some explanation concerning the wall clock time. Finally, we provide an estimation of the carbon footprint required by this paper.

We use the same operator of compression for uplink and downlink, thus we consider that $\omega_{\rm up} = \omega_{\rm dwn}$. In addition, we choose $\alpha_{\rm up} = \alpha_{\rm dwn} = \frac{1}{2(1+\omega_{\rm up/dwn})}$.

Convex settings are given in Table S1. We obtain non-i.i.d. data distributions by computing a TSNE representation [defined in 30] followed by a clustering. Experiments have been performed with 600 epochs. Apart from the case of partial participation, we use quantization [defined in 3] with $s = 2^0$.

Table S1: Settings of experiment in the convex mode.

Settings	a9a	quantum	phishing	superconduct	w8a
references	[10]	[9]	[10]	[16]	[10]
model	LR	LR	LSR	LR	LR
dimension d	124	66	69	82	301
training dataset size	$32,561$	$50,000$	$11,055$	$21,200$	$49,749$
batch size b	50	400	50	50	12
compression rate s	2^0 (*i.e.* two levels)				
norm quantization	$\|\cdot\|_2$				
momentum m	no momentum				
step size γ	$1/L$				

Deep-learning settings are provided in Table S2. All experiments have been performed with 300 epochs

Table S2: Settings of experiments in the non-convex mode.

Settings	MNIST	Fashion-MNIST	FE-MNIST	CIFAR10
references	[26]	[50]	[8]	[24]
model	CNN	Fashion CNN	CNN	LeNet
trainable parameters d	20×10^3	400×10^3	20×10^3	62×10^3
training dataset size	$60,000$	$60,000$	$805,263$	$60,000$
compression rate s	2^2	2^2	2^2	2^4
momentum m	0	0	0	0.9
norm quantization	$\|\cdot\|_2$			
batch size b	128			
step size γ	0.1			
loss	Cross Entropy			

B.1 Convex settings

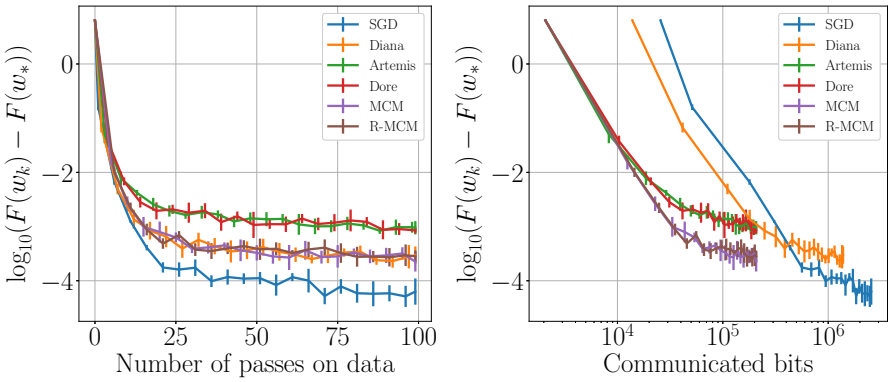
In this section, we provide the plot of excess loss for the toy dataset, for quantum and for a9a datasets. For results on superconduct, phishing and w8a, see our github repository. For these last three datasets, we give only the excess loss w.r.t. number of iteration in the basic settings of full participation on Figure S5. We detail experiments in the PP settings in Appendix B.1.1. At the left side (resp. right side) we display the result w.r.t. the number of iterations (resp. number of communicated bits).

We provide results on the log of the excess loss $F(w_k) - F_*$, with error bars displayed on each figure, corresponding to the standard deviation of $\log_{10}(F(w_k) - F_*)$. Figures S1b, S2b, S3 and S4 correspond to Figures 2a, 2b and 3 given in Section 5. Additionally, we provide results for the synthetic dataset (Figures 2a and 2b) w.r.t to the number of iterations in Figure S1 (stochastic gradient) and Figure S2 (full batch gradient). As predicted by Theorem 2, when $\sigma = 0$, we observe a linear convergence.

On Figure S6, we present a9a, quantum and w8a with a different operator of compression than in all other experiments. We use random unbiased sparsification: each coordinate has a likelihood $p = 0.1$ to be selected.

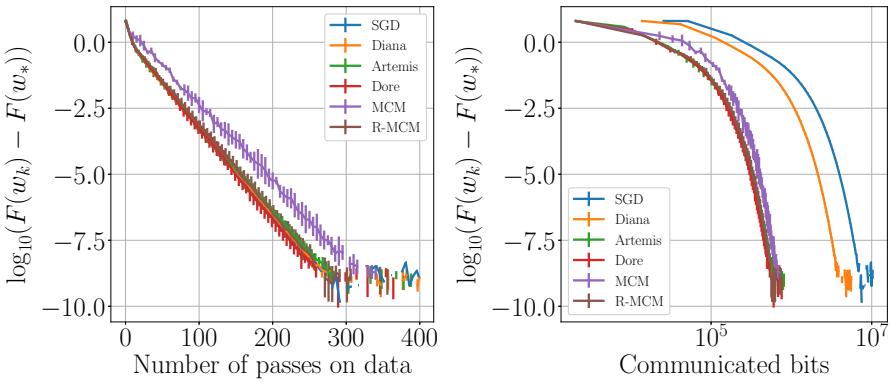
B.1.1 Experiments on partial participation

In this subsection, we run the experiments in a setting where only *half of devices* (independently picked at each iteration) are available at each iteration, thus simulating a setting of partial participation.



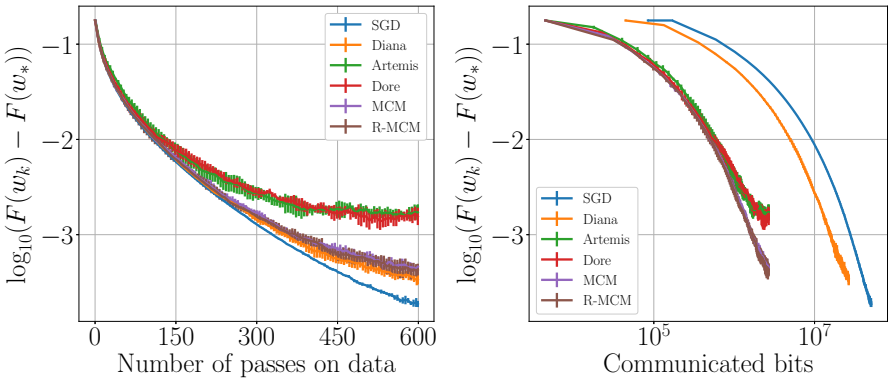
(a) X axis in # iterations. (b) X axis in # bits.

Figure S1: Least-square regression, toy dataset: $\gamma = (L\sqrt{k})^{-1}$, $\sigma \neq 0$.



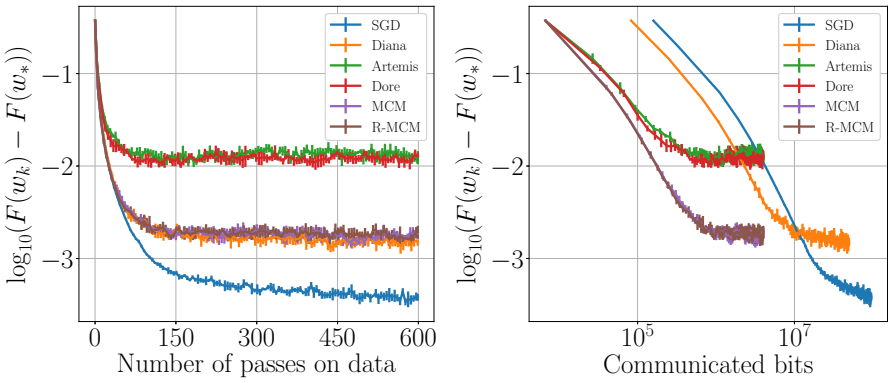
(a) X axis in # iterations. (b) X axis in # bits.

Figure S2: Least-square regression, toy dataset: $\gamma = 1/L$, $\sigma_*^2 = 0$.



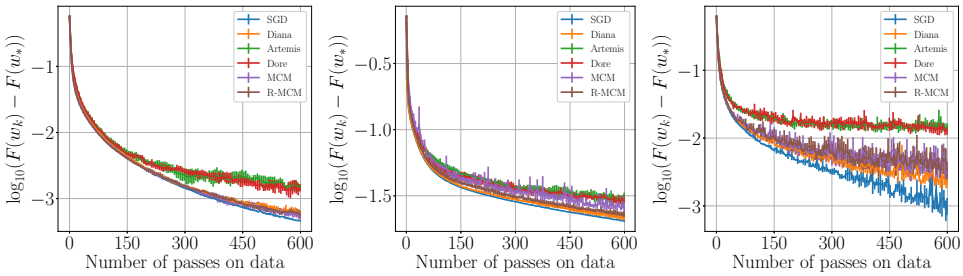
(a) X axis in # iterations. (b) X axis in # bits.

Figure S3: quantum with $b = 400$, $\gamma = 1/L$.



(a) X axis in # iterations. (b) X axis in # bits.

Figure S4: A9A with $b = 50$, $\gamma = 1/L$.



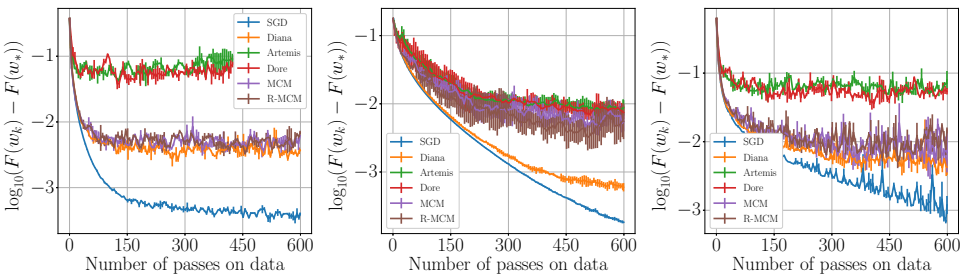
(a) Phishing. (b) Superconduct. (c) W8A.

Figure S5: X axis in # iterations.

Figures S7 and S8 present the results for respectively quantum and A9A. For these experiments, we used a 2-quantization compression. We do not plot `MCM` on these figures because in a context of partial participation, `Rand-MCM` is the natural thing to do. Indeed in this context, we must hold a memory for each worker, and thus the compressed vector sent to each worker is unique.

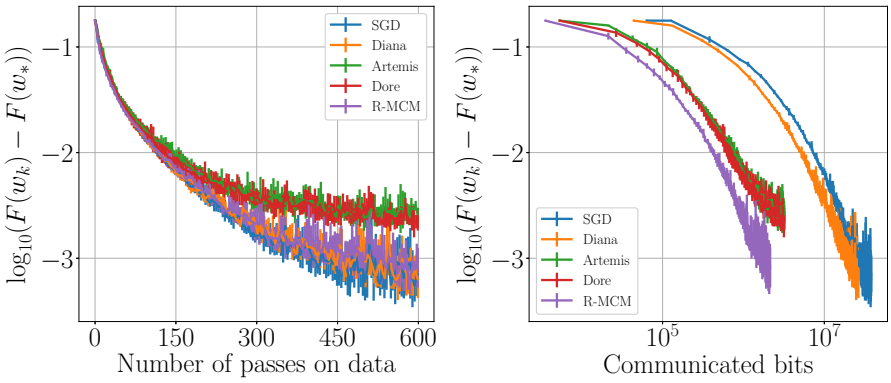
We observe that partial participation leads to an increase of the variance for all algorithms. Furthermore, we can observe on both Figures S7b and S8b that `Rand-MCM` outperforms `Artemis` and `Dore` not only in term of convergence but also in term of communication cost. This is because `Rand-MCM` does not require the synchronization step, at which any active nodes receive any update it has missed. This saves a few communication rounds. In these settings, the level of saturation of `SGD`, `Diana` and `Rand-MCM` seems to be almost identical, this fact stresses again the benefit of our designed algorithm.

Additionally, we present on Figures S9 and S10 the impact of only using a single averaged downlink memory term instead of N distinct memories. More details about update equations are given in Equation (S1). We display three versions of `Rand-MCM` that we compare to the SGD-baseline and to `Artemis`:



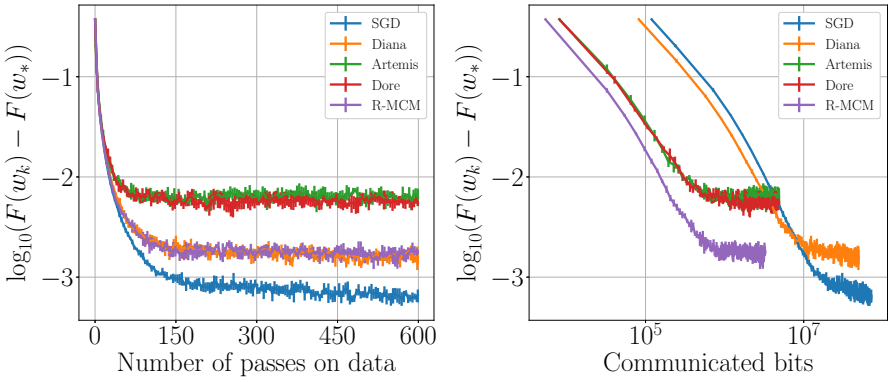
(a) A9A. (b) Quantum. (c) W8A.

Figure S6: X axis in # iterations using random sparsification with $p = 0.1$.



(a) X axis in # iterations. (b) X axis in # bits.

Figure S7: quantum with $b = 400$, $\gamma = 1/L$ and a 2-quantization. Only half of the devices are participating at each round.



(a) X axis in # iterations. (b) X axis in # bits.

Figure S8: A9A with $b = 50$, $\gamma = 1/L$ and a 2-quantization. Only half of the devices are participating at each round.

1. The standard `Rand-MCM`, using N downlink memories,
2. `Rand-MCM` with a single memory, without any periodically reset.
3. `Rand-MCM` with both a single memory and a reset of the downlink memory every $4\sqrt{d}$ iterations, where d is the dimension of the optimization problem. This allows to limit the increase of communicated bits. Indeed as we use quantization with $s = 1$, each communication costs $32 \times \sqrt{d}\log(d)$ bits instead of $32 \times d$. Because every $4\sqrt{d}$ iterations we send the uncompressed downlink memory term, there is an additional cost of $\frac{32d}{4\sqrt{d}}$. At the end, the memory reset leads to send $32 \times \sqrt{d}(\log(d) + 1/4)$ bits by iterations instead of $32 \times \sqrt{d}\log(d)$ bits for `Rand-MCM`(without reset). The increase is thus marginal.

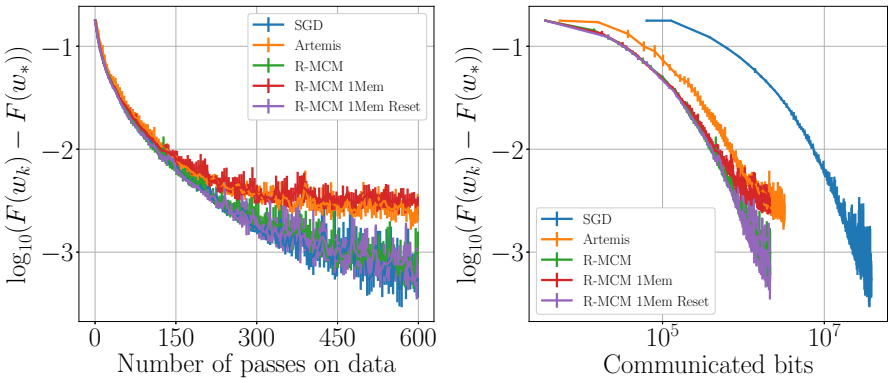
For sake of clarity, we present below the two versions of `Rand-MCM`. In the first version, the central server holds N memories that exactly correspond to those kept on the N remote devices. In the second version, the central server holds a single memory $\bar{H}_k = \frac{1}{N}\sum_{i=1}^N H_k^i$ and each worker i holds there own memory H_k^i.

N memories

$$\begin{cases} \Omega_{k+1}^i = w_{k+1} - H_k^i\,, \\ \widehat{w}_{k+1}^i = H_k^i + \mathcal{C}_{\mathrm{dwn},i}(\Omega_{k+1}^i) \\ H_{k+1}^i = H_k^i + \alpha_{\mathrm{dwn}}\mathcal{C}_{\mathrm{dwn},i}(\Omega_{k+1}^i). \end{cases}$$

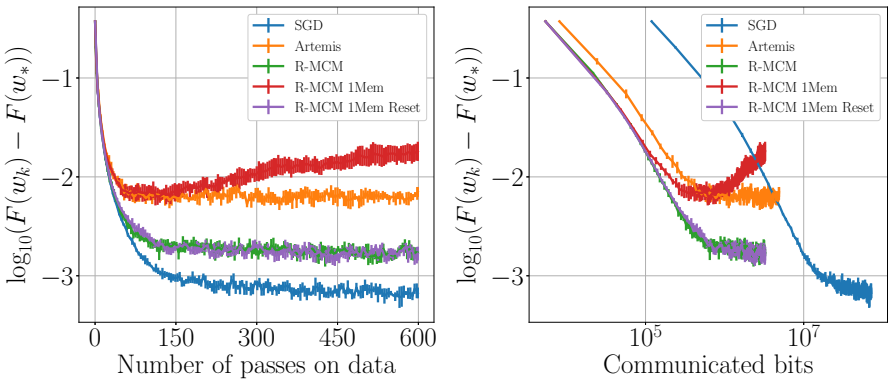
1 memories

$$\begin{cases} \Omega_{k+1} = w_{k+1} - \bar{H}_k\,, \\ \widehat{w}_{k+1}^i = H_k^i + \mathcal{C}_{\mathrm{dwn},i}(\Omega_{k+1}) \\ H_{k+1}^i = H_k^i + \alpha_{\mathrm{dwn}}\mathcal{C}_{\mathrm{dwn},i}(\Omega_{k+1}) \\ \bar{H}_{k+1} = \bar{H}_k + \frac{\alpha_{\mathrm{dwn}}}{N}\sum_{i=1}^N \mathcal{C}_{\mathrm{dwn},i}(\Omega_{k+1}). \end{cases} \tag{S1}$$



(a) X axis in # iterations. (b) X axis in # bits.

Figure S9: quantum with $b = 400$, $\gamma = 1/L$ and a 2-quantization. Only half of the devices are participating at each round.



(a) X axis in # iterations. (b) X axis in # bits.

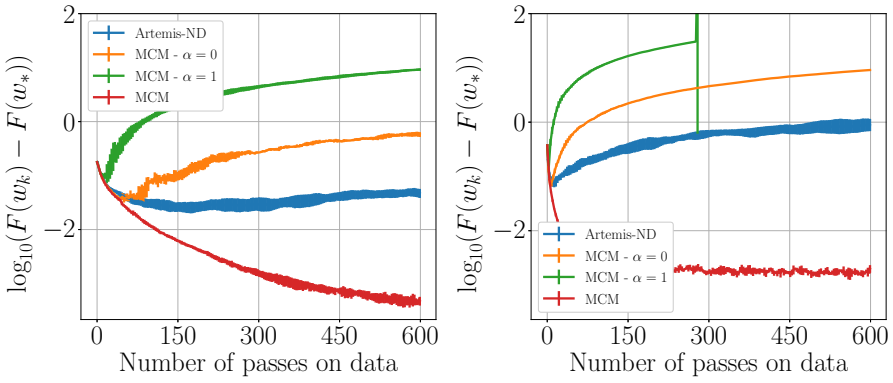
Figure S10: A9A with $b = 50$, $\gamma = 1/L$ and a 2-quantization. Only half of the devices are participating at each round.

In this experiments, it is noticeable that using single-downlink-memory-`Rand-MCM` without periodic reset makes the algorithms saturate at a high level with an important variance. But as soon as we introduce the reset, we recover previous rates.

B.1.2 Comparing `MCM` with other algorithm using non-degraded update

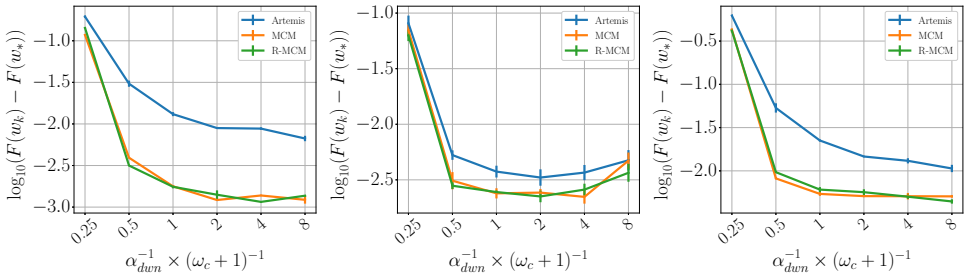
The aim of this section is to show the importance to set $\alpha < 1$, for this purpose we compare `MCM` with three other algorithms:

1. `Artemis` with a non-degraded update i.e. unlike the version proposed by Philippenko and Dieuleveut [36], we do not update the global model with the compression sent to all remote nodes. *It means that we compress only the update that has already been performed on the global server.* It corresponds to:
$$\begin{cases} \forall i \in [\![1, N]\!], \Delta_k^i = g_{k+1}^i(\widehat{w}_k) - h_k^i \\ w_{k+1} = w_k - \frac{\gamma}{N} \sum_{i=1}^N \mathcal{C}_{\text{up}}(\Delta_k^i) + h_k^i \\ \widehat{w}_{k+1} = \widehat{w}_k - \gamma \mathcal{C}_{\text{dwn}} \left(\frac{1}{N} \sum_{i=1}^N \mathcal{C}_{\text{up}}(\Delta_k^i) + h_k^i \right) \\ h_{k+1}^i = h_k^i + \alpha_{\text{up}} \mathcal{C}_{\text{up}}(\Delta_k^i). \end{cases}$$
2. `MCM` with $\alpha = 0$, *thus without memory.*
3. `MCM` with $\alpha = 1$, in other words, for k in $\mathbb{N}^*$ it corresponds to the case $H_{k+1} = \widehat{w}_{k+1}$. Indeed by definition we have $H_{k+1} = H_k + \alpha \widehat{\Omega}_{k+1}$, and furthermore, when we rebuild the compressed model on remote device, we have: $\widehat{w}_{k+1} = \widehat{\Omega}_{k+1} + H_k$. *In this case, we use the compressed model as memory.*



(a) quantum - $b = 400$. (b) *A99* - $b = 50$.

Figure S11: Comparing `MCM` with three other algorithms using a non-degraded update, $\gamma = 1/L$. `Artemis-ND` stands for `Artemis` with a non-degraded update.



(a) A9A. (b) Quantum. (c) W8A.

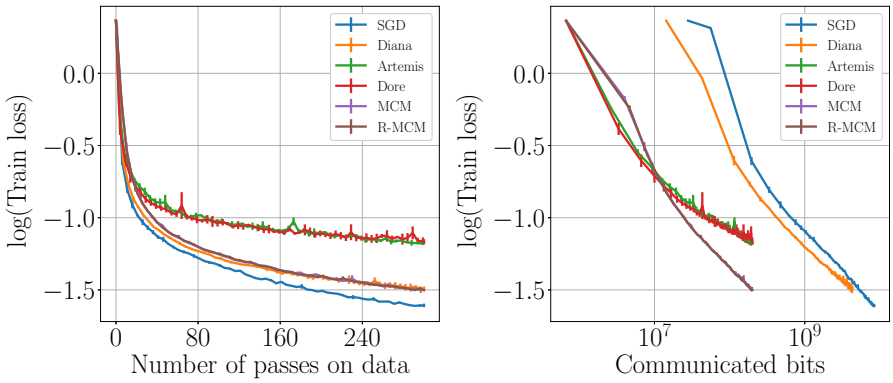
Figure S12: On X axis is displayed different values of $\frac{1}{\alpha(\omega_{\mathrm{dwn}}+1)}$. On Y axis is given the excess loss after 250 epochs. In all other experiments, we choose $\alpha_{\mathrm{dwn}} = \frac{1}{2(\omega_{\mathrm{dwn}}+1)} (= \alpha_{\mathrm{up}})$.

Figures S11a and S11b clearly show the superiority of `MCM` over the three other variants. Some conclusions can be drawn from the observation of these figures.

- `MCM` without downlink memory (orange curve, $\alpha = 0$) does not converge. As stressed in Subsection 2.1, this mechanism is crucial to control the variance of the local model w_{k+1}, for k in $\mathbb{N}$.
- Intuitively, while it appears reasonable to consider as memory the model that has been compressed at the previous step, experiments (green curves) show that this is not the case in practice and that α must be small enough to ensure convergence. This is the *noise explosion* phenomenon that was mentioned earlier in the paper.
- Compressing only the update gives reasonable results (blue curve). However, the convergence saturates at a higher level than for `MCM`.

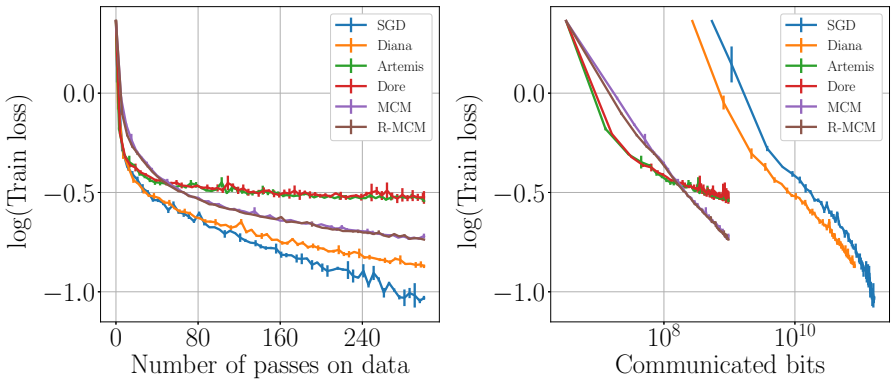
B.1.3 Impact of the learning rate α

On Figure S12, we plot the value of the excess loss obtained after 250 epochs w.r.t. to the value of $\frac{1}{2(1+\omega_{\mathrm{up/dwn}})}$. We observe that if α is too big, `MCM` converges slowly; but after reaching a threshold, the value of α does not impact anymore the rate of convergence. This confirms theory that suggests to use the largest possible α_{dwn} but smaller than a given value. The condition $\alpha_{\mathrm{dwn}} \leq \frac{1}{4(\omega_{\mathrm{dwn}}+1)}$ results from the proofs of Theorems S8 and S14. But because the constant 4 is partially an artifact of the proof, in experiments we used $\alpha_{\mathrm{dwn}} = \frac{1}{2(\omega_{\mathrm{dwn}}+1)}$ as in [36] (see condition S19 in Theorem S7), and this choice is confirmed by Figure S12.



(a) X axis in # iterations. (b) X axis in # bits.

Figure S13: Convergence on MNIST using a CNN.



(a) X axis in # iterations. (b) X axis in # bits.

Figure S14: Convergence on Fashion-MNIST.

B.2 Experiments in deep learning

In this section, we show the robustness of `MCM` in high dimension using more complex data and applying the algorithm to non-convex problems (see Theorem S11 for a guarantee of convergence in this scenario). We carried out experiments on MNIST/FE-MNIST/Fashion-MNIST using a CNN (Figures S13 to S15), and on CIFAR using the LeNet model (Figure S16). We plot the logarithm of the train loss w.r.t the number of iterations and the number of communicated bits. The accuracy has been given in Section 5, see Table 4. Settings of the experiments can be found in Table S2, all experiments are averaged over 2 runs.

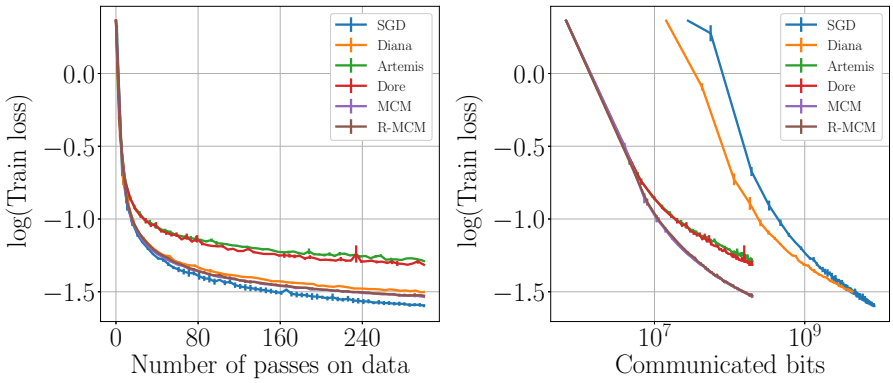
As for experiments in convex case, `MCM` presents identical rates of convergence than `Diana` but with a small shift that makes `Artemis` better during the first iterations.

B.3 Wall clock time

We verified in our experiments that the downlink compression of $w_k - H_{k-1}$ on the central server does not lead to a noticeable overhead w.r.t. gradients computation and communications. Here, as experiments are performed in a *simulated environment* there is no communication cost. In Table S3 we report the computation time when training on FE-MNIST, this allows to highlight that compression only marginally increases the computation cost.

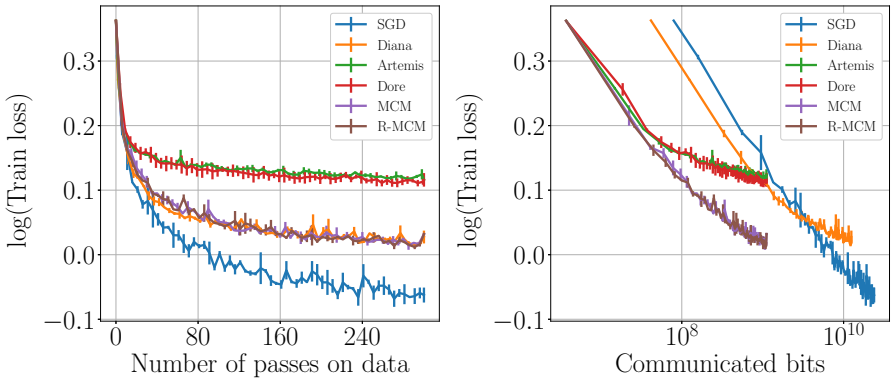
B.4 Hardware and Carbon footprint

As part as a community effort to report the carbon footprint of experiments, we describe in this subsection the hardware used and the total computation time.



(a) X axis in # iterations. (b) X axis in # bits.

Figure S15: Convergence on FE-MNIST.



(a) X axis in # iterations. (b) X axis in # bits.

Figure S16: Convergence on CIFAR10.

Table S3: Wall clock time on FE-MNIST with $b = 128$ and $s = 2^2$.

Compression regime	Computation time for 150 epoch
No compression (SGD)	15421s
Compression on uplink	16773s, ratio: 1.08
Compression on uplink and downlink	16769s, ratio: 1.08

We have two kind of experiments : for deep learning models we ran experiments on a GPU, and for linear/logistic regression on a CPU. We used an Intel(R) Xeon(R) CPU E5-2667 processor with 16 cores; and we used an Nvidia Tesla V100 GPU with 4 nodes.

To generate all figures in this paper, our code ran (if run in a sequential mode) for 150 hours on a CPU. In overall, we consider that the whole paper writing process required (code development, debugging, exploring settings ...) at least 6000 hours end to end on the CPU. The carbon emissions caused by this work were subsequently evaluated with the `Green Algorithm`, built by Lannelongue et al. [25]. It estimates our computations to generate around 100kg of CO2, requiring 2.5MWh. To compare, this corresponds to about 570km by car.

On the GPU, experiments require to be ran for around 140 hours (if run in a sequential mode). In overall, we consider that the full paper writing process required at least 2800 hours end to end on the GPU. The `Green Algorithm` estimates our computations to generate 220kg of CO2, requiring 5.7MWh. To compare, this corresponds to about $1,270$km by car.

C Technical results

In this section, we provide some technical results required by our demonstration. In Appendix C.1 we recall classical inequalities and in Appendix C.2 we present two preliminary lemmas.

In Appendices C to E, for ease of notation we denote, for k in $\mathbb{N}^*$, $\widetilde{g}_k = \frac{1}{N}\sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1})$. Furthermore we use the convention $\nabla F(w_{-1}) = 0$.

C.1 Basic inequalities

In this subsection, we recall some very classical inequalities, for all $a, b \in \mathbb{R}^d$, $\beta > 0$ we have:

$$\langle\, a,\, b\,\rangle \leq \frac{\|a\|^2}{2\beta} + \frac{\beta\,\|b\|^2}{2}\,, \tag{S2}$$

$$\|a+b\|^2 \leq (1+\frac{1}{\beta})\,\|a\|^2 + (1+\beta)\,\|a\|^2\,, \tag{S3}$$

$$\|a+b\|^2 \leq 2\left(\|a\|^2 + \|b\|^2\right)\,, \tag{S4}$$

$$|\,\langle\, a,\, b\,\rangle\,| \leq \|a\|\cdot\|b\| \qquad \text{(Cauchy-Schwarz inequality)}\,, \tag{S5}$$

$$\langle\, a,\, b\,\rangle \leq \frac{1}{2}\left(\|a\|^2 + \|b\|^2 - \|a-b\|^2\right) \qquad \text{(Polarization identity)}. \tag{S6}$$

Below, we recall Jensen's inequality.

Jensen inequality Let a probability space $(\Omega, \mathcal{A}, \mathbf{P})$ with Ω a sample space, $\mathcal{A}$ an event space, and $\mathbf{P}$ a probability measure. Suppose that $X : \Omega \longrightarrow \mathbb{R}^d$ is a random variable, then for any convex function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ we have:

$$f\left(\mathbb{E}(X)\right) \leq \mathbb{E}f(X)\,. \tag{S7}$$

The next lemma will be used several times in the proofs.

Lemma S1. *Let a probability space $(\Omega, \mathcal{A}, \mathbf{P})$ with Ω a sample space, $\mathcal{A}$ an event space, $\mathbf{P}$ a probability measure, and $\mathcal{F}$ a $\sigma-$algebra. For any $a \in \mathbb{R}^d$ and for any random vector in $\mathbb{R}^d$ we have:*

$$\mathbb{E}\left[\|X - \mathbb{E}X\|^2\right] \leq \mathbb{E}\left[\|X-a\|^2\right]$$

indeed $\mathbb{E}[X] = \arg\min_{a\in\mathbb{R}^d}\mathbb{E}\left[\|X-a\|^2\right]$. Similarly, for any random vector Y in $\mathbb{R}^d$ which is $\mathcal{F}$-measurable, we have:

$$\mathbb{E}\left[\|X - \mathbb{E}\left[X \mid \mathcal{F}\right]\|^2 \;\middle|\; \mathcal{F}\right] \leq \mathbb{E}\left[\|X-Y\|^2 \;\middle|\; \mathcal{F}\right]\,.$$

Assumption 5 (Cocoercivity)**.** *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:*

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

This assumption is stronger than supposing convexity and L-smoothness of F.

The final proposition of this subsection presents two inequalities used in our demonstrations when invoking convexity or strong-convexity. They follow from Assumption 3 and can be found in [35].

Proposition S1. *If a function F is convex, then it satisfies for all w in $\mathbb{R}^d$:*

$$\langle\, \nabla F(x),\; w - w_*\,\rangle \geq \frac{1}{2}(F(w) - F(w_*)) + \frac{1}{2L}\,\|\nabla F(w)\|^2\,. \tag{S8}$$

If a function F is strongly-convex, then it satisfies for all w in $\mathbb{R}^d$:

$$\langle\, \nabla F(x),\; w - w_*\,\rangle \geq \frac{1}{2}(F(w) - F(w_*)) + \frac{1}{2}\left(\mu\,\|w-w_*\|^2 + \frac{1}{L}\,\|\nabla F(w)\|^2\right)\,. \tag{S9}$$

C.2 Two lemmas

In this subsection, we give two lemmas required to prove the convergences of `Ghost`[3], `MCM` and `Rand-MCM`.

The first lemma will be used to show that `MCM` indeed satisfies Theorem 3. The proof is straightforward from the definition of w_k and H_{k-1}.

Lemma S2 (Expectation of $w_k - H_{k-1}$)**.** *For any k in $\mathbb{N}^*$, the expectation of $(w_k - H_{k-1})$ conditionally to w_{k-1} can be decomposed as follows:*

$$\mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right] = (1-\alpha_{\mathrm{dwn}})(w_{k-1} - H_{k-2}) - \gamma \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right] .$$

Proof. Let k in $\mathbb{N}^*$, by definition and with Assumption 1:

$$\begin{aligned}\mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right] &= \mathbb{E}\left[w_{k-1} - \gamma \widehat{g}_k(\widehat{w}_{k-1}) - (H_{k-2} + \alpha_{\mathrm{dwn}}\mathcal{C}(w_{k-1} - H_{k-2})) \mid w_{k-1}\right] \\ &= (w_{k-1} - H_{k-2}) - \alpha_{\mathrm{dwn}}\mathbb{E}\left[\mathcal{C}\left(w_{k-1} - H_{k-2}\right) \mid w_{k-1}\right] - \gamma \mathbb{E}\left[\widetilde{g}_k \mid w_{k-1}\right] ,\end{aligned}$$

from which the result follows.

□

The following lemma provides a control of the impact of the uplink compression. It decomposes the squared-norm of stochastic gradients into two terms: 1) the true gradient 2) the variance of the stochastic gradient σ^2.

Lemma S3 (Squared-norm of stochastic gradients)**.** *For any k in $\mathbb{N}^*$, the second moment and variance of the compressed gradients can be bounded a.s.:*

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{g}_k\|^2 \;\middle|\; \widehat{w}_{k-1}\right] &\leq \left(1 + \frac{\omega_{\mathrm{up}}}{N}\right)\|\nabla F(\widehat{w}_{k-1})\|^2 + \frac{\sigma^2(1+\omega_{\mathrm{up}})}{Nb} , \\ \mathbb{E}\left[\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; \widehat{w}_{k-1}\right] &\leq \frac{\omega_{\mathrm{up}}}{N}\|\nabla F(\widehat{w}_{k-1})\|^2 + \frac{\sigma^2(1+\omega_{\mathrm{up}})}{Nb} .\end{aligned}$$

Interpretation:

- *If $\omega_{\mathrm{up}} = 0$ (i.e. no up compression), the variance corresponds to a mini-batch.*
- *If $\sigma = 0$ and $N = 1$ (i.e. full batch descent with a single device), it becomes: $\mathbb{E}\left[\|\mathcal{C}(\nabla F(w_{k-1})) - \nabla F(w_{k-1})\|^2\right] \leq \omega_{\mathrm{up}}\|\nabla F(w_{k-1})\|^2$ which is consistent with Assumption 1.*

Proof. Let k in $\mathbb{N}^*$, then $\mathbb{E}\left[\|\widetilde{g}_k\|^2 \;\middle|\; \widehat{w}_{k-1}\right] = \|\nabla F(\widehat{w}_{k-1})\|^2 + \mathbb{E}\left[\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; \widehat{w}_{k-1}\right]$.

Secondly:

$$\begin{aligned}&\mathbb{E}\left[\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; \widehat{w}_{k-1}\right] \\ &\quad= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \left(\widehat{g}_k^i(\widehat{w}_{k-1}) - \nabla F(\widehat{w}_{k-1})\right)\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right] \\ &\quad= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \left(\widehat{g}_k^i(\widehat{w}_{k-1}) - g_k^i(\widehat{w}_{k-1}) + g_k^i(\widehat{w}_{k-1}) - \nabla F(\widehat{w}_{k-1})\right)\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right] \\ &\quad= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \left(\widehat{g}_k^i(\widehat{w}_{k-1}) - g_k^i(\widehat{w}_{k-1})\right)\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right] \\ &\qquad+ \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \left(g_k^i(\widehat{w}_{k-1}) - \nabla F(\widehat{w}_{k-1})\right)\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right] ,\end{aligned}$$

[3] `Ghost` is defined in Appendix D.1.

the inner product being null.

Next expanding the squared norm again, and because the two sums of inner products are null as the stochastic oracle and uplink compressions are independent:

$$\mathbb{E}\left[\left\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\right\|^2 \,\middle|\, \widehat{w}_{k-1}\right] = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|\widehat{g}_k^i(\widehat{w}_{k-1}) - g_k^i(\widehat{w}_{k-1})\right\|^2 \,\middle|\, \widehat{w}_{k-1}\right] + \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i(\widehat{w}_{k-1}) - \nabla F(\widehat{w}_{k-1})\right\|^2 \,\middle|\, \widehat{w}_{k-1}\right].$$

Then, for any i in $[\![1, N]\!]$ as $\mathbb{E}\left[\left\|\widehat{g}_k^i(\widehat{w}_{k-1}) - g_k^i(\widehat{w}_{k-1})\right\|^2 \,\middle|\, \widehat{w}_{k-1}\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|\widehat{g}_k^i(\widehat{w}_{k-1}) - g_k^i(\widehat{w}_{k-1})\right\|^2 \,\middle|\, g_k^i\right] \,\middle|\, \widehat{w}_{k-1}\right]$, and using Assumption 1 we have:

$$\mathbb{E}\left[\left\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\right\|^2 \,\middle|\, \widehat{w}_{k-1}\right] = \frac{\omega_{\rm up}}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i(\widehat{w}_{k-1})\right\|^2 \,\middle|\, \widehat{w}_{k-1}\right] + \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i(\widehat{w}_{k-1}) - \nabla F(\widehat{w}_{k-1})\right\|^2 \,\middle|\, \widehat{w}_{k-1}\right].$$

Furthermore $\mathbb{E}\left[\left\|g_k^i(\widehat{w}_{k-1})\right\|^2 \,\middle|\, \widehat{w}_{k-1}\right] = \mathbb{E}\left[\left\|g_k^i(\widehat{w}_{k-1}) - \nabla F(\widehat{w}_{k-1})\right\|^2 \,\middle|\, \widehat{w}_{k-1}\right] + \left\|\nabla F(\widehat{w}_{k-1})\right\|^2$, and using Assumption 4:

$$\mathbb{E}\left[\left\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\right\|^2 \,\middle|\, \widehat{w}_{k-1}\right] = \frac{\omega_{\rm up}}{N}\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 + \frac{\sigma^2(1+\omega_{\rm up})}{Nb},$$

from which we derive the two inequalities of the lemma.

□

D The `Ghost` algorithm

D.1 Motivation, definition of `Ghost` and proof sketch

In this section, to convey the best understanding of the theorems and the spirit of the proof, we define a *ghost* algorithm (that is impossible to implement in practice). `Ghost` is introduced only to get some intuition of the theoretical insight.

Definition 1 (`Ghost` algorithm)**.** *The `Ghost` algorithm is defined as follows, for $k \in \mathbb{N}$, for all $i \in [\![1, N]\!]$ we have:*

$$w_{k+1} = w_k - \gamma \frac{1}{N}\sum_{i=1}^N \widehat{g}_{k+1}^i(\widehat{w}_k) \quad \textit{and} \quad \widehat{w}_{k+1} = w_k - \gamma \mathcal{C}_{\rm dwn}\left(\frac{1}{N}\sum_{i=1}^N \widehat{g}_{k+1}^i(\widehat{w}_k)\right). \tag{S10}$$

While the global model is unchanged (1^{st} line), the local model $\widehat{w}_{k+1}$ (2^{nd} line) is updated using the global model w_k at the previous step, which is not available locally.

In the following, we give the main results for `Ghost` and complete them with a sketch of proof. Demonstrations are all in the next subsection.

The following Proposition, provides the control of the variance of the local model for `Ghost`.

Proposition S2. *Consider the `Ghost` update in eq. (S10), under Assumptions 1, 2 and 4, for all k in $\mathbb{N}$ with the convention $\nabla F(w_{-1}) = 0$:*

$$\mathbb{E}\left[\left\|w_k - \widehat{w}_k\right\|^2 \,\middle|\, \widehat{w}_{k-1}\right] \le \gamma^2 \omega_{\rm dwn}\left(1 + \frac{\omega_{\rm up}}{N}\right)\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 + \frac{\gamma^2 \omega_{\rm dwn}(1+\omega_{\rm up})\sigma^2}{Nb}.$$

Proof. The proof of Proposition S2 is straightforward using Definition 1. Let k in $\mathbb{N}$, by Definition 1 we have:

$$\begin{aligned}\|w_k - \widehat{w}_k\|^2 &= \left\| \left(w_{k-1} - \gamma \mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}) \right) \right) - \left(w_{k-1} - \gamma \frac{1}{N} \sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}) \right) \right\|^2 \\ &= \gamma^2 \left\| \mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}) \right) - \frac{1}{N} \sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}) \right\|^2 .\end{aligned}$$

Taking expectation w.r.t. down compression, as $\frac{1}{N} \sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1})$ is w_k-measurable:

$$\mathbb{E}\left[\|w_k - \widehat{w}_k\|^2 \;\middle|\; w_k \right] = \gamma^2 \omega_{\mathrm{dwn}} \mathbb{E}\left[\left\| \frac{1}{N} \sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}) \right\|^2 \;\middle|\; w_k \right] = \gamma^2 \omega_{\mathrm{dwn}} \|\widetilde{g}_k\|^2 ,$$

and Lemma S3 gives the upper bound $\mathbb{E}\left[\|\widetilde{g}_k\|^2 \;\middle|\; \widehat{w}_{k-1} \right]$. □

The takeaway from this Proposition is that we are able to bound the variance of the local model by an affine function of the squared norm of the *previous* stochastic gradients $\nabla F(\widehat{w}_{k-1})$. For `Ghost` only the previous gradient is involved, while for `MCM`, we obtain an additional recursive process.

To obtain the convergence, we then follow the classical approach [31], expanding $\mathbb{E}\|w_k - w_*\|^2$ as $\mathbb{E}\|w_{k-1} - w_*\|^2 - 2\gamma \mathbb{E} \langle \nabla F(\widehat{w}_{k-1}),\, w_{k-1} - w_* \rangle + \gamma^2 \mathbb{E}\left[\|\widehat{g}_k(\widehat{w}_{k-1})\|^2 \right]$. The critical aspect is that the inner product does not directly result in a contraction, as the support point of the gradient differs from w_{k-1}. Using the fact that $\mathbb{E}\left[\widehat{w}_{k-1} \mid w_{k-1} \right] = w_{k-1}$, we further decompose it as

$$-2\gamma \mathbb{E} \langle \nabla F(\widehat{w}_{k-1}),\, \widehat{w}_{k-1} - w_* \rangle + 2\gamma \mathbb{E} \langle \nabla F(\widehat{w}_{k-1}) - \nabla F(w_{k-1}),\, w_{k-1} - \widehat{w}_{k-1} \rangle . \quad \text{(S11)}$$

The first part of eq. (S11), corresponds to a "strong contraction": by (strong-)convexity, we can upper bound it by $-2\gamma(\mu \|\widehat{w}_{k-1} - w_*\|^2 + F(\widehat{w}_{k-1}) - F_*)$, which is on average larger than $-2\gamma(\mu \|w_{k-1} - w_*\|^2 + F(w_{k-1}) - F_*)$ (Jensen's inequality). Moreover, as the function is smooth and convex, it can also be upper bounded by $-2\gamma \|\nabla F(\widehat{w}_{k-1})\|^2 / L$. This is a crucial term: we "gain" something of the order of a squared norm of the gradient at $\widehat{w}_{k-1}$, which will *in fine* compensate the variance of the local model. The second part of eq. (S11), corresponds to a positive residual term, proportional to the variance of the compressed model, that can be controlled thanks to Proposition S2 (at w_{k-1}!). Putting things together, we get, in the convex case ($\mu = 0$):

Theorem S6 (Contraction for `Ghost`, convex case)**.** *Under Assumptions 1 to 4, with $\mu = 0$, if $\gamma L(1 + \omega_{\mathrm{up}}/N) \leq \frac{1}{2}$.*

$$\begin{aligned}\mathbb{E}\|w_k - w_*\|^2 \leq \mathbb{E}\|w_{k-1} - w_*\|^2 - \gamma \mathbb{E}(F(w_{k-1}) - F_*) - \frac{\gamma}{2L} \mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2 \right] \\ + 2\gamma^3 \omega_{\mathrm{dwn}} L \left(1 + \frac{\omega_{\mathrm{up}}}{N} \right) \mathbb{E}\|\nabla F(\hat{w}_{k-2})\|^2 + \gamma^2 \frac{(1 + \omega_{\mathrm{up}})\sigma^2}{Nb} \left(1 + 2\gamma L \omega_{\mathrm{dwn}} \right) .\end{aligned}$$

We can make the following observations:

1. At step k, the residual can be upper bounded by a constant times squared norm of the gradient at point $\hat{w}_{k-2}$. When using recursively this upper bound, if $2\gamma^3 \omega_{\mathrm{dwn}} L(1 + \omega_{\mathrm{up}}/N) \leq \gamma/(2L)$, then these terms cancel out. This is equivalent to $2\gamma L \sqrt{\omega_{\mathrm{dwn}} (1 + \omega_{\mathrm{up}}/N)} \leq 1$. It is natural to chose $\gamma \leq 1/(2L \max(1 + \omega_{\mathrm{up}}/N, 1 + \omega_{\mathrm{dwn}}))$.
2. The bound is in fact proved conditionally to w_{k-1}, recursive conditioning is required to propagate the inequality. We carefully handle conditioning in the proofs.

D.2 Convergence of `Ghost`, complete proof

In this subsection, we provide the complete proof of convergence for `Ghost`. Thus in the following demonstration, we give the key concepts required to later prove the convergence of `MCM`.

Theorem S7 (Convergence of `Ghost`, convex case)**.** *Under Assumptions 1 to 4 with $\mu = 0$ (convex case), for all k in $\mathbb{N}$, defining $V_k := \mathbb{E}\left[w_k - w_*\right] + \frac{\gamma}{2L}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right] + 2\gamma L\mathbb{E}\left[\left\|\widehat{w}_k - w_k\right\|^2\right]$, we have:*

$$V_k \leq V_{k-1} - \gamma\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] + \frac{\gamma^2\sigma^2\Phi^{\mathcal{G}}(\gamma)}{Nb}\,,$$

with $\Phi^{\mathcal{G}}(\gamma) := (1+\omega_{\rm up})(1+2\gamma L\omega_{\rm dwn})$.

Remark 6. *This result is similar to eq. (6) but with a different function $\Phi^{\mathcal{G}}$ that has a weaker dependency on $\omega_{\rm dwn}$.*

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

$$\left\|w_k - w_*\right\|^2 \leq \left\|w_{k-1} - w_*\right\|^2 - 2\gamma\left\langle\, \widetilde{g}_k,\, w_{k-1} - w_*\,\right\rangle + \gamma^2\left\|\widetilde{g}_k\right\|^2\,.$$

Next, we expend the inner product as following:

$$\left\|w_k - w_*\right\|^2 \leq \left\|w_{k-1} - w_*\right\|^2 - 2\gamma\left\langle\, \widetilde{g}_k,\, \widehat{w}_{k-1} - w_*\,\right\rangle - 2\gamma\left\langle\, \widetilde{g}_k,\, w_{k-1} - \widehat{w}_{k-1}\,\right\rangle + \gamma^2\left\|\widetilde{g}_k\right\|^2\,.$$

Taking expectation conditionally to w_{k-1}, and using $\mathbb{E}\left[\widetilde{g}_k \mid w_{k-1}\right] = \mathbb{E}\left[\mathbb{E}\left[\widetilde{g}_k \mid \widehat{w}_{k-1}\right] \mid w_{k-1}\right] = \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]$, we obtain:

$$\begin{aligned}\mathbb{E}\left[\left\|w_k - w_*\right\|^2 \,\middle|\, w_{k-1}\right] \leq \left\|w_{k-1} - w_*\right\|^2 &- \mathbb{E}\left[2\gamma\left\langle\, \nabla F(\widehat{w}_{k-1}),\, \widehat{w}_{k-1} - w_*\,\right\rangle \mid w_{k-1}\right] \\ &- 2\gamma\mathbb{E}\left[\left\langle\, \nabla F(\widehat{w}_{k-1}),\, w_{k-1} - \widehat{w}_{k-1}\,\right\rangle \mid w_{k-1}\right] \\ &+ \gamma^2\mathbb{E}\left[\left\|\widetilde{g}_k\right\|^2 \,\middle|\, w_{k-1}\right]\,.\end{aligned}$$

Then invoking Lemma S3 to upper bound the squared norm of the stochastic gradients, and noticing that $\mathbb{E}\left[\left\langle\, \nabla F(w_{k-1}),\, \widehat{w}_{k-1} - w_{k-1}\,\right\rangle \mid w_{k-1}\right] = 0$ leads to:

$$\begin{aligned}\mathbb{E}\left[\left\|w_k - w_*\right\|^2 \,\middle|\, w_{k-1}\right] \leq \left\|w_{k-1} - w_*\right\|^2 &- 2\gamma\mathbb{E}\left[\left\langle\, \nabla F(\widehat{w}_{k-1}),\, \widehat{w}_{k-1} - w_*\,\right\rangle \mid w_{k-1}\right] \\ &- 2\gamma\mathbb{E}\left[\left\langle\, \nabla F(\widehat{w}_{k-1}) - \nabla F(w_{k-1}),\, w_{k-1} - \widehat{w}_{k-1}\,\right\rangle \mid w_{k-1}\right] \\ &+ \gamma^2\left(\left(1 + \frac{\omega_{\rm up}}{Nb}\right)\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \,\middle|\, w_{k-1}\right] + \frac{\sigma^2\left(1+\omega_{\rm up}\right)}{Nb}\right)\,.\end{aligned} \tag{S12}$$

In the upper inequality:

1. the term $\mathbb{E}\left[\left\langle\, \nabla F(\widehat{w}_{k-1}),\, \widehat{w}_{k-1} - w_*\,\right\rangle \mid w_{k-1}\right]$ allows the "strong contraction"
2. the terms $\mathbb{E}\left[\left\langle\, \nabla F(\widehat{w}_{k-1}) - \nabla F(w_{k-1}),\, w_{k-1} - \widehat{w}_{k-1}\,\right\rangle \mid w_{k-1}\right]$ and $\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \,\middle|\, w_{k-1}\right]$ are two positives terms that we treat as residuals.
3. the last term $\sigma^2\left(1+\omega_{\rm up}\right)/(Nb)$ is due to the stochastic noise.

Now using Cauchy-Schwarz inequality (eq. (S5)) and smoothness:

$$\begin{aligned}&-\mathbb{E}\left[2\gamma\left\langle\, \nabla F(\widehat{w}_{k-1}) - \nabla F(w_{k-1}),\, w_{k-1} - \widehat{w}_{k-1}\,\right\rangle \mid w_{k-1}\right] \\ &\quad= 2\gamma\mathbb{E}\left[\left\langle\, \nabla F(\widehat{w}_{k-1}) - \nabla F(w_{k-1}),\, \widehat{w}_{k-1} - w_{k-1}\,\right\rangle \mid w_{k-1}\right] \\ &\quad\leq 2\gamma L\mathbb{E}\left[\left\|\widehat{w}_{k-1} - w_{k-1}\right\|^2 \,\middle|\, w_{k-1}\right]\,,\end{aligned}$$

and thus:

$$\begin{aligned}\mathbb{E}\left[\left\|w_k - w_*\right\|^2 \,\middle|\, w_{k-1}\right] \leq \left\|w_{k-1} - w_*\right\|^2 &- 2\gamma\mathbb{E}\left[\left\langle\, \nabla F(\widehat{w}_{k-1}),\, \widehat{w}_{k-1} - w_*\,\right\rangle \mid w_{k-1}\right] \\ &+ 2\gamma L\mathbb{E}\left[\left\|\widehat{w}_{k-1} - w_{k-1}\right\|^2 \,\middle|\, w_{k-1}\right] \\ &+ \gamma^2\left(1 + \frac{\omega_{\rm up}}{N}\right)\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \,\middle|\, w_{k-1}\right] + \frac{\gamma^2\sigma^2(1+\omega_{\rm up})}{Nb}\,.\end{aligned} \tag{S13}$$

Now, using convexity with Proposition S1:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2 \;\middle|\; w_{k-1}\right] \leq \|w_{k-1} - w_*\|^2 \\ - \gamma\mathbb{E}\left[\left(F(\widehat{w}_{k-1}) - F(w_*) + \frac{1}{L}\|\nabla F(\widehat{w}_{k-1})\|^2\right) \;\middle|\; w_{k-1}\right] \\ + 2\gamma L\mathbb{E}\left[\|\widehat{w}_{k-1} - w_{k-1}\|^2 \;\middle|\; w_{k-1}\right] \\ + \gamma^2\left(1 + \frac{\omega_{\text{up}}}{N}\right)\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; w_{k-1}\right] + \frac{\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb}\,.\end{aligned}$$

Taking the full expectation (without conditioning over any random vectors), and because invoking Jensen inequality (S7) leads to $\mathbb{E}\left[F(\widehat{w}_{k-1})\right] \geq \mathbb{E}\left[F(w_{k-1})\right]$, we finally obtain this intermediate result:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2\right] \leq \mathbb{E}\left[\|w_{k-1} - w_*\|^2\right] - \gamma\left(\mathbb{E}\left[F(w_{k-1})\right] - F(w_*)\right) \\ - \frac{\gamma}{2L}\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] \\ + 2\gamma L\mathbb{E}\left[\|\widehat{w}_{k-1} - w_{k-1}\|^2\right] + \frac{\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb}\,,\end{aligned} \tag{S14}$$

where we considered that $\gamma L(1 + \omega_{\text{up}}/N) \leq 1/2$, which implies that $\gamma\left(1 - \gamma L\left(1 + \frac{\omega_{\text{up}}}{N}\right)\right) \geq \frac{\gamma}{2}$.

Remark that eq. (S14) is valid for both `Ghost` and `MCM`, and that the proof of `MCM` will follow the same initial line.

With Proposition S2:

$$\mathbb{E}\left[\|w_k - \widehat{w}_k\|^2 \;\middle|\; \widehat{w}_{k-1}\right] \leq \gamma^2\omega_{\text{dwn}}\left(1 + \frac{\omega_{\text{up}}}{N}\right)\|\nabla F(\widehat{w}_{k-1})\|^2 + \frac{\gamma^2\omega_{\text{dwn}}(1+\omega_{\text{up}})\sigma^2}{Nb}\,. \tag{S15}$$

Defining $V_k := \mathbb{E}\left[w_k - w_*\right] + \frac{\gamma}{2L}\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] + C\mathbb{E}\left[\|\widehat{w}_k - w_k\|^2\right]$ with $C = 2\gamma L$, and combining this two equations as following $(S14) + C(S15)$ leads to:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2\right] + C\mathbb{E}\left[\|\widehat{w}_{k-1} - w_{k-1}\|^2\right] + \frac{\gamma}{2L}\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] \\ \leq \mathbb{E}\left[\|w_{k-1} - w_*\|^2\right] - \gamma\left(\mathbb{E}\left[F(w_{k-1})\right] - F(w_*)\right) \\ + 2\gamma L\mathbb{E}\left[\|\widehat{w}_{k-1} - w_{k-1}\|^2\right] + \frac{\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb} \\ + 2\gamma L \times \gamma^2\omega_{\text{dwn}}\left(1 + \frac{\omega_{\text{up}}}{N}\right)\|\nabla F(\widehat{w}_{k-1})\|^2 + 2\gamma L \times \frac{\gamma^2\omega_{\text{dwn}}(1+\omega_{\text{up}})\sigma^2}{Nb}\,.\end{aligned}$$

To ensure a contraction of the Lyapunov function we require:

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

Under this condition, we obtain:

$$V_k \leq V_{k-1} - \gamma\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] + \frac{\gamma^2\sigma^2\Phi^{\mathcal{G}}(\gamma)}{Nb}\,,$$

with $\Phi^{\mathcal{G}}(\gamma) := (1 + \omega_{\text{up}})(1 + 2\gamma L\omega_{\text{dwn}})$.

By recurrence and for $k = K$:

$$V_K \leq V_0 - \sum_{k=1}^{K}\gamma\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] + \sum_{k=1}^{K}\frac{\gamma^2\sigma^2\Phi^{\mathcal{G}}(\gamma)}{Nb}\,,$$

which leads to:

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

Finally, for any K in $\mathbb{N}^*$, with $\gamma L\leq\min\left\{\frac{1}{2\left(1+\frac{\omega_{\rm up}}{N}\right)},\frac{1}{2\sqrt{\omega_{\rm dwn}\left(1+\frac{\omega_{\rm up}}{N}\right)}}\right\}$ we have:

$$\frac{\gamma}{K}\sum_{t=1}^{K}\mathbb{E}\left[F(w_t)-F(w_*)\right]\leq\frac{\|w_0-w_*\|^2}{K}+\frac{\gamma\sigma^2\Phi^{\mathcal{G}}(\gamma)}{Nb}\,.$$

Note that the bound of γL encompass the case $\omega_{\rm dwn}=0$ (i.e. no downlink compression), but in the general case of bidirectional compression, we nearly always have $\omega_{\rm dwn}>1$, and thus the dominant term is in fact $\frac{1}{2\sqrt{\omega_{\rm dwn}\left(1+\frac{\omega_{\rm up}}{N}\right)}}$.

And by Jensen, it implies that:

$$\mathbb{E}\left[F(\bar{w}_K)-F(w_*)\right]\leq\frac{\|w_0-w_*\|^2}{\gamma K}+\frac{\gamma\sigma^2\Phi(\gamma)}{Nb}\quad\text{with }\Phi^{\mathcal{G}}(\gamma):=(1+\omega_{\rm up})(1+2\omega_{\rm dwn}\gamma L)\,.$$

□

E Proofs for `MCM` (and `Rand-MCM`)

In this section, we provide the proofs for `MCM` in the convex, strongly-convex, and non-convex cases in respectively Theorems S9 to S11. The proofs for `Rand-MCM` (see Theorem 4) are identical and only require to adapt notations as explained in appendix E.5.

We denote for γ in $\mathbb{R}$, $\Phi(\gamma)\;:=\;(1+\omega_{\rm up})\left(1+\frac{8\gamma L\omega_{\rm dwn}}{\alpha_{\rm dwn}}\right)$, for k in $\mathbb{N}$, $\Upsilon_k=\|w_k-H_{k-1}\|^2$ and we define $\gamma_{\max}$ such that:

$$\gamma_{\max}L\leq\min\left\{\frac{1}{8\omega_{\rm dwn}},\frac{1}{2\left(1+\frac{\omega_{\rm up}}{N}\right)},\frac{1}{4\sqrt{\frac{\omega_{\rm dwn}}{\alpha_{\rm dwn}}\left(\frac{1}{\alpha_{\rm dwn}}+\frac{\omega_{\rm up}}{N}\right)}}\right\}.$$

Note that this is equivalent to notations given in Section 3 if we take $\alpha_{\rm dwn}=1/8\omega_{\rm dwn}$.

E.1 Control of the Variance of the local model for `MCM` (Theorem 3)

In this section, we provide a control of the variance of the local model for `MCM`, as done previously in Proposition S2 for `Ghost`: this corresponds to Theorem 3. The demonstration is more complex than for `Ghost` and it highlights the trade-offs for the learning rate $\alpha_{\rm dwn}$. The demonstration builds a bias-variance decomposition of $\|\Omega_k\|^2=\|w_k-H_k\|^2$. The variance is then decomposed in three terms, as a result we will need to compute four terms:

$$\|w_k-H_{k-1}\|^2=\text{Bias}^2+2\gamma^2(\text{Var}_{11}+\text{Var}_{12})+2\alpha_{\rm dwn}^2\text{Var}_2\,.\tag{S16}$$

Theorem S8. *Consider the `MCM` update as in eq. (2). Under Assumptions 1, 2 and 4 with $\mu=0$, if $\gamma\leq(8\omega_{\rm dwn}L)^{-1}$ and $\alpha_{\rm dwn}\leq(8\omega_{\rm dwn})^{-1}$, then for all k in $\mathbb{N}$:*

$$\mathbb{E}\left[\Upsilon_k\right]\leq\left(1-\frac{\alpha_{\rm dwn}}{2}\right)\mathbb{E}\left[\Upsilon_{k-1}\right]+2\gamma^2\left(\frac{1}{\alpha_{\rm dwn}}+\frac{\omega_{\rm up}}{N}\right)\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right]+\frac{2\gamma^2\sigma^2(1+\omega_{\rm up})}{Nb}\,.$$

Proof. Let k in $\mathbb{N}$, we recall that by definition:

$$\begin{cases} \Omega_k = w_k - H_{k-1} \\ \widehat{\Omega}_k = \mathcal{C}_{\mathrm{dwn}}(\Omega_k) \\ \widehat{w}_k = \widehat{\Omega}_k + H_{k-1} \,. \end{cases}$$

We start the proof by introducing $\|\Omega_k\|^2$:

$$\mathbb{E}\left[\|w_k - \widehat{w}_k\|^2 \;\middle|\; w_k\right] = \mathbb{E}\left[\left\|\widehat{\Omega}_k - \Omega_k\right\|^2 \;\middle|\; w_k\right] \leq \omega_{\mathrm{dwn}} \|\Omega_k\|^2 \,.$$

Next, we perform a bias-variance decomposition:

$$\begin{aligned} \|\Omega_k\|^2 = \|w_k - H_{k-1}\|^2 = \|w_k - H_{k-1} - \mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right]\|^2 \\ + \|\mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right]\|^2 \\ + 2\left\langle w_k - H_{k-1} - \mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right], \mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right] \right\rangle \,, \end{aligned}$$

taking expectation w.r.t. w_{k-1}:

$$\begin{aligned} \mathbb{E}\left[\Upsilon_k \mid w_{k-1}\right] = \underbrace{\mathbb{E}\left[\|w_k - H_{k-1} - \mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right]\|^2 \;\middle|\; w_{k-1}\right]}_{\mathrm{Var}} \\ + \underbrace{\|\mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right]\|^2}_{\mathrm{Bias}^2} \,. \end{aligned}$$

The first term is the variance Var, and the second term corresponds to the squared bias Bias^2.

Let's handle first the variance, by definition:

$$\begin{aligned} \mathrm{Var} &= \mathbb{E}\left[\|w_k - H_{k-1} - \mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right]\|^2 \;\middle|\; w_{k-1}\right] \\ &= \mathbb{E}\left[\|w_{k-1} - \gamma\widetilde{g}_k - H_{k-2} - \alpha_{\mathrm{dwn}}\mathcal{C}(w_{k-1} - H_{k-2}) \right. \\ &\qquad \left. - w_{k-1} - \gamma\mathbb{E}\left[\widetilde{g}_k \mid w_{k-1}\right] - H_{k-2} - \alpha_{\mathrm{dwn}}\mathbb{E}\left[\mathcal{C}(w_{k-1} - H_{k-2} \mid w_{k-1})\right)\|^2 \middle| w_{k-1}\right] \,. \end{aligned}$$

After simplification and using eq. (S4):

$$\begin{aligned} \mathrm{Var} &= \mathbb{E}\left[\| - \gamma\left(\widetilde{g}_k + \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right) + \alpha_{\mathrm{dwn}}\left(\mathcal{C}(w_{k-1} - H_{k-2})\right) \right. \\ &\qquad \left. - (w_{k-1} - H_{k-2})\|^2 \middle| w_{k-1}\right] \\ &\leq 2\gamma^2\mathbb{E}\left[\|\widetilde{g}_k - \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\|^2 \;\middle|\; w_{k-1}\right] \\ &\qquad + 2\alpha_{\mathrm{dwn}}^2\mathbb{E}\left[\|\mathcal{C}(w_{k-1} - H_{k-2}) - (w_{k-1} - H_{k-2})\|^2 \;\middle|\; w_{k-1}\right] \\ &\leq 2\gamma^2 \underbrace{\mathbb{E}\left[\|\widetilde{g}_k - \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\|^2 \;\middle|\; w_{k-1}\right]}_{\mathrm{Var}_1} + 2\alpha_{\mathrm{dwn}}^2 \underbrace{\omega_{\mathrm{dwn}} \|w_{k-1} - H_{k-2}\|^2}_{\mathrm{Var}_2} \\ &\leq 2\gamma^2\mathrm{Var}_1 + 2\alpha_{\mathrm{dwn}}^2\mathrm{Var}_2 \,. \end{aligned}$$

An interpretation of the above decomposition is that:

- Var_1 is the part of the downlink compression caused by the increment $\widetilde{g}_k$, it is similar to `Ghost`.
- Var_2 is the impact of the propagation of the previous noise.

We compute the first term by introducing $\nabla F(\widehat{w}_{k-1})$, the second being kept as it is:

$$\begin{aligned} \mathrm{Var}_1 &= \mathbb{E}\left[\|\widetilde{g}_k - \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\|^2 \;\middle|\; w_{k-1}\right] \\ &= \mathbb{E}\left[\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1}) + \nabla F(\widehat{w}_{k-1}) - \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\|^2 \;\middle|\; w_{k-1}\right] \\ &= \underbrace{\mathbb{E}\left[\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; w_{k-1}\right]}_{\mathrm{Var}_{11}} + \underbrace{\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1}) - \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\|^2 \;\middle|\; w_{k-1}\right]}_{\mathrm{Var}_{12}} \\ &= \mathrm{Var}_{11} + \mathrm{Var}_{12} \,, \end{aligned}$$

the inner product is null given that $\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) - \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right] \mid w_{k-1}\right] = 0$.
Moreover:

$$\text{Var}_{11} = \mathbb{E}\left[\left\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right] \;\middle|\; w_{k-1}\right],$$

so, we can use Lemma S3: $\text{Var}_{11} = \mathbb{E}\left[\frac{\sigma^2}{Nb}(1+\omega_{\text{up}}) + \frac{\omega_{\text{up}}}{N}\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right]$.

And now we use smoothness for the second term:

$$\begin{aligned}
\text{Var}_{12} &= \mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1}) - \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right\|^2 \;\middle|\; w_{k-1}\right] \\
&\leq \mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1}) - \nabla F(w_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right] \quad \text{by Lemma S1,} \\
&\leq L^2\mathbb{E}\left[\left\|\widehat{w}_{k-1} - w_{k-1}\right\|^2 \;\middle|\; w_{k-1}\right] \quad \text{using smoothness,} \\
&\leq L^2\omega_{\text{dwn}}\Upsilon_{k-1} \quad \text{with Assumption 1 .}
\end{aligned}$$

At the end:

$$\begin{aligned}
\text{Var} \leq 2\gamma^2 &\left(\frac{\sigma^2(1+\omega_{\text{up}})}{Nb} + \frac{\omega_{\text{up}}}{N}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right] + L^2\omega_{\text{dwn}}\Upsilon_{k-1}\right) \\
&+ 2\alpha_{\text{dwn}}^2\omega_{\text{dwn}}\Upsilon_{k-1}\,.
\end{aligned} \tag{S17}$$

Now we focus on the squared bias Bias^2, with Lemma S2:

$$\begin{aligned}
\text{Bias}^2 &= \left\|\mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right]\right\|^2 \\
&= \left\|(1-\alpha_{\text{dwn}})(w_{k-1} - H_{k-2}) - \gamma\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right\|^2, \quad \text{and with Equation (S3),} \\
&\leq (1-\alpha_{\text{dwn}})^2(1+\alpha_{\text{dwn}})\,\Upsilon_{k-1} + \gamma^2\left(1+\frac{1}{\alpha_{\text{dwn}}}\right)\left\|\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right\|^2 .
\end{aligned}$$

And because $(1-\alpha_{\text{dwn}})(1+\alpha_{\text{dwn}}) < 1$, we finally get that:

$$\text{Bias}^2 \leq (1-\alpha_{\text{dwn}})\Upsilon_{k-1} + \gamma^2\left(1+\frac{1}{\alpha_{\text{dwn}}}\right)\left\|\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right\|^2 . \tag{S18}$$

Combining all eqs. (S17) and (S18) into eq. (S16):

$$\begin{aligned}
\mathbb{E}\left[\Upsilon_k \mid w_{k-1}\right] \leq\; & (1-\alpha_{\text{dwn}})\Upsilon_{k-1} + \gamma^2\left(1+\frac{1}{\alpha_{\text{dwn}}}\right)\left\|\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right\|^2 \\
&+ 2\gamma^2\left(\frac{\sigma^2(1+\omega_{\text{up}})}{Nb} + \frac{\omega_{\text{up}}}{N}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right]\right) \\
&+ 2\gamma^2\left(L^2\omega_{\text{dwn}}\Upsilon_{k-1}\right) \\
&+ 2\alpha_{\text{dwn}}^2\omega_{\text{dwn}}\Upsilon_{k-1} \\
\leq\; & \left(1-\alpha_{\text{dwn}} + 2\gamma^2L^2\omega_{\text{dwn}} + 2\alpha_{\text{dwn}}^2\omega_{\text{dwn}}\right)\left\|w_{k-1} - H_{k-2}\right\|^2 \\
&+ \gamma^2\left(1+\frac{1}{\alpha_{\text{dwn}}}\right)\left\|\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right\|^2 \\
&+ \frac{2\gamma^2\omega_{\text{up}}}{N}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right] + \frac{2\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb}\,.
\end{aligned}$$

Next, we require:

$$\begin{cases}
2\alpha_{\text{dwn}}^2\omega_{\text{dwn}} \leq \frac{1}{4}\alpha_{\text{dwn}} \iff \alpha_{\text{dwn}} \leq \frac{1}{8\omega_{\text{dwn}}}\,, \\
2\gamma^2L^2\omega_{\text{dwn}} \leq \frac{1}{4}\alpha_{\text{dwn}} = \frac{1}{32\omega_{\text{dwn}}}\,, \text{ by taking } \alpha_{\text{dwn}} = \frac{1}{8\omega_{\text{dwn}}} \quad \iff \gamma \leq \frac{1}{8\omega_{\text{dwn}}L}\,, \\
1 + \frac{1}{\alpha_{\text{dwn}}} \leq \frac{2}{\alpha_{\text{dwn}}} \text{ which is not restrictive if } \omega_{\text{dwn}} \geq 1.
\end{cases}$$

Thus, it leads to:

$$\mathbb{E}\left[\Upsilon_k \mid w_{k-1}\right] \leq \left(1 - \frac{\alpha_{\text{dwn}}}{2}\right)\Upsilon_{k-1} + \frac{2\gamma^2}{\alpha_{\text{dwn}}}\left\|\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right\|^2 + \frac{2\gamma^2\omega_{\text{up}}}{N}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right] + \frac{2\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb}\,.$$

Next, we bound $\left\|\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right\|^2$ with $\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right]$, and we obtain:

$$\mathbb{E}\left[\Upsilon_k \mid w_{k-1}\right] \leq \left(1 - \frac{\alpha_{\text{dwn}}}{2}\right)\Upsilon_{k-1} + 2\gamma^2\left(\frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N}\right)\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right] + \frac{2\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb}\,.$$

Taking the unconditional expectation gives the result.

□

E.2 Convex case (Theorem 2)

In this section, we give the demonstration of MCM in the convex case (Theorem 2).

Theorem S9 (Convergence of MCM in the homogeneous and convex case)**.** *Under Assumptions 1 to 4 with $\mu = 0$, for a learning rate $\alpha_{\text{dwn}} \leq \frac{1}{8\omega_{\text{dwn}}}$, for all $k > 0$, for any $\gamma \leq \gamma_{\max}$, defining $V_k := \mathbb{E}[\|w_k - w_*\|^2] + 32\gamma L\omega_{\text{dwn}}^2\mathbb{E}[\Upsilon_k]$, for $\bar{w}_k = \frac{1}{k}\sum_{i=0}^{k-1} w_i$, we have:*

$$\gamma\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] \leq V_{k-1} - V_k + \frac{\gamma^2\sigma^2\Phi(\gamma)}{Nb} \Longrightarrow \mathbb{E}[F(\bar{w}_k) - F_*] \leq \frac{V_0}{\gamma k} + \frac{\gamma\sigma^2\Phi(\gamma)}{Nb}\,.$$

Consequently, for K in $\mathbb{N}$ large enough, a step-size $\gamma = \sqrt{\frac{\|w_0 - w_*\|^2 Nb}{(1+\omega_{\text{up}})\sigma^2 K}}$ and a learning rate $\alpha_{\text{dwn}} = \frac{1}{8\omega_{\text{dwn}}}$, we have:

$$\mathbb{E}[F(\bar{w}_K) - F_*] \leq 2\sqrt{\frac{\|w_0 - w_*\|^2(1+\omega_{\text{up}})\sigma^2}{NbK}} + O(K^{-1}).$$

Moreover if $\sigma^2 = 0$ (noiseless case), we recover a faster convergence: $\mathbb{E}[F(\bar{w}_K) - F_*] = O(K^{-1})$.

Proof. Let k in $\mathbb{N}^*$, the proof follows the one for Ghost, and we start from eq. (S14):

$$\mathbb{E}\left[\|w_k - w_*\|^2\right] \leq \mathbb{E}\left[\|w_{k-1} - w_*\|^2\right] - \gamma\left(\mathbb{E}\left[F(w_{k-1})\right] - F(w_*)\right) - \frac{\gamma}{2L}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right] + 2\gamma L\mathbb{E}\left[\left\|\widehat{w}_{k-1} - w_{k-1}\right\|^2\right] + \frac{\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb}\,,$$

with Assumption 1, it easily becomes:

$$\mathbb{E}\left[\|w_k - w_*\|^2\right] \leq \mathbb{E}\left[\|w_{k-1} - w_*\|^2\right] - \gamma\left(\mathbb{E}\left[F(w_{k-1})\right] - F(w_*)\right) - \frac{\gamma}{2L}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right] + 2\gamma L\omega_{\text{dwn}}\mathbb{E}\left[\Upsilon_{k-1}\right] + \frac{\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb}\,.$$

Theorem 3 which is specific to MCM gives:

$$\mathbb{E}\left[\Upsilon_k\right] \leq \left(1 - \frac{\alpha_{\text{dwn}}}{2}\right)\mathbb{E}\left[\Upsilon_{k-1}\right] + 2\gamma^2\left(\frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N}\right)\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right] + \frac{2\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb}\,.$$

Defining: $V_k := \mathbb{E}\left[\|w_k - w_*\|^2\right] + \gamma LC\mathbb{E}\left[\Upsilon_k\right]$ with $C = \frac{4\omega_{\rm dwn}}{\alpha_{\rm dwn}}$, and, combining the two last equations:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2\right] + \gamma LC\mathbb{E}\left[\Upsilon_k\right] \leq \mathbb{E}\left[\|w_{k-1} - w_*\|^2\right] &- \gamma\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] \\ &+ 2\gamma L\omega_{\rm dwn}\mathbb{E}\left[\Upsilon_{k-1}\right] \\ &- \frac{\gamma}{2L}\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] + \frac{\gamma^2\sigma^2(1+\omega_{\rm up})}{Nb} \\ &+ \left(1 - \frac{\alpha_{\rm dwn}}{2}\right)\gamma LC\mathbb{E}\left[\Upsilon_{k-1}\right] \\ &+ 2\gamma^3 LC\left(\frac{1}{\alpha_{\rm dwn}} + \frac{\omega_{\rm up}}{N}\right)\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] \\ &+ \frac{2\gamma^3 L\sigma^2(1+\omega_{\rm up})C}{N}\,,\end{aligned}$$

and reordering the terms gives:

$$\begin{aligned}V_k \leq \mathbb{E}\left[\|w_{k-1} - w_*\|^2\right] &+ \left(2\gamma L\omega_{\rm dwn} + \left(1 - \frac{\alpha_{\rm dwn}}{2}\right)\gamma LC\right)\mathbb{E}\left[\|w_{k-1} - H_{k-1}\|^2\right] \\ &+ \left(2\gamma^3 LC\left(\frac{1}{\alpha_{\rm dwn}} + \frac{\omega_{\rm up}}{N}\right) - \frac{\gamma}{2L}\right)\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] \\ &- \gamma\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] \\ &+ (2\gamma LC + 1)\frac{\gamma^2\sigma^2(1+\omega_{\rm up})}{Nb}\,.\end{aligned}$$

We observe that:

$$2\gamma L\omega_{\rm dwn} + \left(1 - \frac{\alpha_{\rm dwn}}{2}\right)\gamma LC \leq \gamma LC \iff C \geq \frac{4\omega_{\rm dwn}}{\alpha_{\rm dwn}} \quad \text{which is true by definition of } C.$$

Secondly, to get the contraction requires

$$\begin{aligned}2\gamma^3 LC\left(\frac{1}{\alpha_{\rm dwn}} + \frac{\omega_{\rm up}}{N}\right) - \frac{\gamma}{2L} \leq 0 &\iff \gamma^2 L \leq \frac{1}{4LC\left(\frac{1}{\alpha_{\rm dwn}} + \frac{\omega_{\rm up}}{N}\right)} \\ &\iff \gamma L \leq \frac{1}{4\sqrt{\frac{\omega_{\rm dwn}}{\alpha_{\rm dwn}}\left(\frac{1}{\alpha_{\rm dwn}} + \frac{\omega_{\rm up}}{N}\right)}}\,,\end{aligned}$$

because $C = 4\omega_{\rm dwn}/\alpha$. Thus, we have that:

$$V_k \leq V_{k-1} - \gamma\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] + \frac{\gamma^2\sigma^2\Phi(\gamma)}{Nb} \quad \text{denoting } \Phi(\gamma) := (1+\omega_{\rm up})\left(1 + \frac{8\gamma L\omega_{\rm dwn}}{\alpha_{\rm dwn}}\right),$$

and then for $k = K \in \mathbb{N}^*$, by recurrence:

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

which implies:

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

Finally, by Jensen, for any K in $\mathbb{N}^*$ such that $\gamma L \leq \min\left\{\frac{1}{8\omega_{\rm dwn}}, \frac{1}{2\left(1+\frac{\omega_{\rm up}}{N}\right)}, \frac{1}{4\sqrt{\frac{\omega_{\rm dwn}}{\alpha_{\rm dwn}}\left(\frac{1}{\alpha_{\rm dwn}}+\frac{\omega_{\rm up}}{N}\right)}}\right\}$, we have:

$$\mathbb{E}\left[F(\bar{w}_K) - F(w_*)\right] \leq \frac{V_0}{\gamma K} + \frac{\gamma\sigma^2\Phi(\gamma)}{Nb}\,,$$

which concludes the proof.

□

E.3 Strongly-convex case (Theorem 1)

In this section, we give the demonstration for MCM in the strongly-convex case (Theorem 1).

Theorem S10 (Convergence of MCM in the homogeneous and strongly-convex case)**.** *Under Assumptions 1 to 4 with $\mu > 0$, for k in $\mathbb{N}$, for a learning rate $\alpha_{\rm dwn} \leq \frac{1}{8\omega_{\rm dwn}}$, for any sequence $(\gamma_k)_{k\geq 0} \leq \gamma_{\max}$, defining $V_k := \mathbb{E}[\|w_k - w_*\|^2] + 32\gamma L\omega_{\rm dwn}^2\mathbb{E}[\Upsilon_k]$, we have:*

$$V_k \leq (1-\gamma_k\mu)V_{k-1} - \gamma_k\mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] + \frac{\gamma_k^2\sigma^2\Phi(\gamma_k)}{Nb}\,,$$

Consequently,

1. *if $\sigma^2 = 0$ (noiseless case), for $\gamma_k \equiv \gamma_{\max}$ we recover a linear convergence rate: $\mathbb{E}[\|w_K) - w_*\|^2] \leq (1-\gamma_{\max}\mu)^k V_0$;*
2. *if $\sigma^2 > 0$, defining $\widetilde{L}$ such that $\gamma_{\max} = (2\widetilde{L})^{-1}$, taking for all k in $\mathbb{N}$, $\gamma_k = 2/(\mu(k+1)+\widetilde{L})$, for the weighted Polyak-Ruppert average $\bar{w}_K = \sum_{k=1}^K \lambda_k w_{k-1}/\sum_{k=1}^K \lambda_k$, with $\lambda_k := \frac{1}{\gamma_{k-1}^{-1}}$, we have:*

$$\mathbb{E}\left[F(\bar{w}_K) - F(w_*)\right] \leq \frac{\mu+2\widetilde{L}}{4\mu K^2}\|w_0 - w_*\|^2 + \frac{4\sigma^2(1+\omega_{\rm up})}{\mu K N b}\left(1 + \frac{64L\omega_{\rm dwn}^2}{\mu K}\ln(\mu K + \widetilde{L})\right).$$

Proof. Let k in $\mathbb{N}^*$, the proof starts like the one for Ghost, and we start from eq. (S13) but we consider a variable step size $\gamma_k = 2/(\mu(k+1)+\widetilde{L})$ that depends of the iteration k in $\mathbb{N}$.

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2\right] \leq \mathbb{E}\left[\|w_{k-1} - w_*\|^2\right] &- 2\gamma_k\mathbb{E}\left[\langle \nabla F(\widehat{w}_{k-1}),\, \widehat{w}_{k-1} - w_* \rangle\right] \\ &+ 2\gamma_k L\mathbb{E}\left[\|\widehat{w}_{k-1} - w_{k-1}\|^2\right] + \gamma_k^2\left(1+\frac{\omega_{\rm up}}{N}\right)\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] \\ &+ \frac{\gamma_k^2\sigma^2(1+\omega_{\rm up})}{Nb}\,.\end{aligned}$$

Now we apply strong-convexity (eq. (S9) of Proposition S1):

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2\right] \leq \mathbb{E}\left[\|w_{k-1} - w_*\|^2\right] &+ 2\gamma_k L\mathbb{E}\left[\|\widehat{w}_{k-1} - w_{k-1}\|^2\right] \\ &- \gamma_k\mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] - \gamma_k\left(\mu\|\widehat{w}_{k-1} - w_*\|^2 + \frac{1}{L}\|\nabla F(\widehat{w}_{k-1})\|^2\right) \\ &+ \gamma_k^2\left(1+\frac{\omega_{\rm up}}{N}\right)\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1})\|^2\right] + \frac{\gamma_k^2\sigma^2(1+\omega_{\rm up})}{Nb}\,.\end{aligned}$$

As $\gamma_k \leq \frac{2}{\widetilde{L}} \leq \frac{1}{2L\left(1+\frac{\omega_{\mathrm{up}}}{N}\right)}$, and thus $\left(1-\gamma_k L\left(1+\frac{\omega_{\mathrm{up}}}{N}\right)\right) \geq 1/2$; this allows to simplify the coefficient of $\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right]$:

$$\begin{aligned}\mathbb{E}\left[\left\|w_k - w_*\right\|^2\right] \leq (1-\gamma_k\mu)\left\|w_{k-1}-w_*\right\|^2 - \gamma_k\mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] \\ - \frac{\gamma_k}{2L}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right] + 2\gamma_k L\mathbb{E}\left[\left\|\widehat{w}_{k-1} - w_{k-1}\right\|^2\right] \\ + \frac{\gamma_k^2\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\end{aligned}$$

equivalent to:

$$\begin{aligned}\mathbb{E}\left[\left\|w_k - w_*\right\|^2\right] \leq (1-\gamma_k\mu)\left\|w_{k-1}-w_*\right\|^2 - \gamma_k\mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] \\ - \frac{\gamma_k}{2L}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right] + 2\gamma_k L\omega_{\mathrm{dwn}}\mathbb{E}\left[\left\|w_{k-1} - H_{k-1}\right\|^2\right] \\ + \frac{\gamma_k^2\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\,.\end{aligned} \tag{S19}$$

Theorem 3 adapted to the case of decaying steps gives:

$$\begin{aligned}\mathbb{E}\left[\Upsilon_k\right] \leq \left(1-\frac{\alpha_{\mathrm{dwn}}}{2}\right)\mathbb{E}\left[\Upsilon_{k-1}\right] + 2\gamma_k^2\left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{\omega_{\mathrm{up}}}{N}\right)\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right] \\ + \frac{2\gamma_k^2\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\,.\end{aligned} \tag{S20}$$

Defining $V_k := \mathbb{E}\left[\left\|w_k - w_*\right\|^2\right] + \gamma_k LC\mathbb{E}\left[\Upsilon_k\right]$ with $C = 4\omega_{\mathrm{dwn}}/\alpha$, combining the two later equations (S9) + $\gamma_k LC$ (S20):

$$\begin{aligned}&\mathbb{E}\left[\left\|w_k - w_*\right\|^2\right] + \gamma_k LC\mathbb{E}\left[\Upsilon_k\right] \\ &\quad\leq (1-\gamma_k\mu)\left\|w_{k-1}-w_*\right\|^2 - \gamma_k\mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] \\ &\qquad - \frac{\gamma_k}{2L}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right] + 2\gamma_k L\omega_{\mathrm{dwn}}\mathbb{E}\left[\left\|w_{k-1} - H_{k-1}\right\|^2\right] + \frac{\gamma_k^2\sigma^2(1+\omega_{\mathrm{up}})}{Nb} \\ &\qquad + \left(1-\frac{\alpha_{\mathrm{dwn}}}{2}\right)\gamma_k LC\mathbb{E}\left[\Upsilon_{k-1}\right] + 2\gamma_k^3 LC\left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{\omega_{\mathrm{up}}}{N}\right)\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right] \\ &\qquad + \frac{2\gamma_k^3 L\sigma^2(1+\omega_{\mathrm{up}})C}{Nb}\,,\end{aligned}$$

and reordering the terms gives:

$$\begin{aligned}V_k \leq (1-\gamma_k\mu)\left\|w_{k-1}-w_*\right\|^2 - \gamma_k\mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] \\ + \left(1-\frac{\alpha_{\mathrm{dwn}}}{2} + \frac{2\omega_{\mathrm{dwn}}}{C}\right)\gamma_k LC\mathbb{E}\left[\left\|w_{k-1} - H_{k-1}\right\|^2\right] \\ + \left(2\gamma_k^3 LC\left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{\omega_{\mathrm{up}}}{N}\right) - \frac{\gamma_k}{2L}\right)\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right] \\ + (2\gamma_k LC + 1)\,\frac{\gamma_k^2\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\,,\end{aligned}$$

To reach a $(1-\gamma\mu)$-convergence we first need $\left(1-\frac{\alpha_{\mathrm{dwn}}}{2} + \frac{2\omega_{\mathrm{dwn}}}{C}\right)\gamma_k LC \leq (1-\gamma_k\mu)\gamma_{k-1}LC$ i.e $1-\frac{\alpha_{\mathrm{dwn}}}{2} + \frac{2\omega_{\mathrm{dwn}}}{C} \leq \frac{(1-\gamma_k\mu)\gamma_{k-1}}{\gamma_k}$.

We need that for all $k \in \mathbb{N}$, $\frac{1-\gamma_k\mu}{\gamma_k} \leq \frac{1}{\gamma_{k-1}}$ i.e., $1-\gamma_k\mu \leq \frac{\gamma_k}{\gamma_{k-1}}$, but:

$$\frac{\gamma_k}{\gamma_{k-1}} = \frac{\mu k - \mu + \widetilde{L}}{\mu k + \widetilde{L}} = 1 - \frac{\mu}{\mu k + \widetilde{L}} \quad \text{and} \quad 1 - \gamma_k\mu = 1 - \frac{2\mu}{\mu k + \widetilde{L}}\,,$$

and so, the inequality is always true.

Thus we must have $2\omega_{\mathrm{dwn}}/C \leq \alpha_{\mathrm{dwn}}/2$ which is true by definition of C.

Secondly, it requires:

$$\begin{aligned} 2\gamma_k^2 C\left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{\omega_{\mathrm{up}}}{N}\right) - \frac{\gamma_k}{2L} \leq 0 &\iff \gamma_k L \leq \frac{1}{4C\left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{\omega_{\mathrm{up}}}{N}\right)} \\ &\iff \gamma_k L \leq \frac{1}{4\sqrt{\frac{\omega_{\mathrm{dwn}}}{\alpha_{\mathrm{dwn}}}\left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{\omega_{\mathrm{up}}}{N}\right)}}\,, \end{aligned}$$

by definition of C. And it follows that the first part of the theorem is proved:

$$V_k \leq (1-\gamma_k\mu)V_{k-1} - \gamma_k\mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] + \frac{\gamma_k^2\sigma^2\Phi(\gamma_k)}{Nb}\,,$$

where $\Phi(\gamma_k) := (1+\omega_{\mathrm{up}})\left(1 + \frac{8\gamma_k L\omega_{\mathrm{dwn}}}{\alpha_{\mathrm{dwn}}}\right)$.

We now prove the second part, which requires to carefully handle the term of noise. By definition $\gamma_k = \frac{2}{\mu(k+1)+L}$, we denote $\lambda_k = \frac{1}{\gamma_{k-1}}$ and we sum the above equation weighted with the sequence of $(\lambda_k)_{k=1}^K$:

$$\begin{aligned} \frac{1}{\sum_{k=1}^K \lambda_k}\sum_{k=1}^K \lambda_k\mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] \leq \frac{1}{\sum_{k=1}^K \lambda_k}\sum_{k=1}^K \frac{(1-\gamma_k\mu)\lambda_k}{\gamma_k}V_{k-1} - \frac{\lambda_k}{\gamma_k}V_k \\ + \frac{1}{\sum_{k=1}^K \lambda_k}\sum_{k=1}^K \lambda_k\frac{\gamma_k\sigma^2\Phi(\gamma_k)}{Nb}\,. \end{aligned}$$

The weights are chosen to ensure that the sum of $(V_k)_{k=1}^K$ is telescopic. Because $(1-\gamma_k\mu)/\gamma_k = \gamma_{k-2}^{-1}$, we have:

$$\begin{aligned} \frac{1}{\sum_{k=1}^K \lambda_k}\sum_{k=1}^K \lambda_k\mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] \leq \frac{1}{\sum_{k=1}^K \lambda_k}\sum_{k=1}^K \frac{1}{\gamma_{k-2}\gamma_{k-1}}V_{k-1} - \frac{1}{\gamma_k\gamma_{k-1}}V_k \\ + \frac{1}{\sum_{k=1}^K \lambda_k}\sum_{k=1}^K \lambda_k\frac{\gamma_k\sigma^2\Phi(\gamma_k)}{Nb}\,, \end{aligned}$$

and because for $K \in \mathbb{N}^*$ big enough $\frac{1}{\sum_{k=1}^K \lambda_k} = \frac{1}{\mu(K+1)K/4 + (\widetilde{L}K)/2} \leq \frac{4}{\mu K^2}$, it results that:

$$\frac{1}{\sum_{k=1}^K \lambda_k}\sum_{k=1}^K \lambda_k\mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] \leq \frac{V_0}{\gamma_0\gamma_{-1}\mu K^2} + \frac{4}{\mu K^2}\sum_{k=1}^K \lambda_k\frac{\gamma_k\sigma^2\Phi(\gamma_k)}{Nb}\,. \tag{S21}$$

At the end, using the Jensen inequality - $\mathbb{E}\left[\mathbb{E}\left[F(\widehat{w}_{k-1}) \mid w_{k-1}\right]\right] \leq \mathbb{E}\left[F(w_{k-1})\right]$, see Equation (S7) - we have for all K in $\mathbb{N}$:

$$\begin{aligned}
&\frac{1}{\sum_{k=1}^{K}\lambda_k}\sum_{k=1}^{K}\lambda_k\mathbb{E}\left[F(w_{k-1})-F(w_*)\right] \\
&\qquad\leq \frac{V_0}{\gamma_0\gamma_{-1}\mu K^2}+\frac{4}{\mu K^2}\sum_{k=1}^{K}\frac{1}{\gamma_{k-1}}\left(1+\frac{8\gamma_k L\omega_{\rm dwn}}{\alpha_{\rm dwn}}\right)\frac{\gamma_k\sigma^2(1+\omega_{\rm up})}{Nb} \\
&\qquad\leq \frac{V_0}{\gamma_0\gamma_{-1}\mu K^2}+\frac{4}{\mu K^2}\sum_{k=1}^{K}\left(1+\frac{8\gamma_k L\omega_{\rm dwn}}{\alpha_{\rm dwn}}\right)\frac{\sigma^2(1+\omega_{\rm up})}{Nb}\,,
\end{aligned}$$

because for all k in N^*, $\gamma_k \leq \gamma_{k-1}$. We need to compute the following classical sum:

$$\sum_{k=1}^{K}\frac{1}{\mu k+\widetilde{L}} \leq \int_{x=0}^{K}\frac{1}{\mu x+\widetilde{L}}\mathrm{dx} \leq \frac{1}{\mu}\ln\left(\mu\mathrm{K}+\tilde{\mathrm{L}}\right)\,.$$

At the end, using again the Jensen inequality, defining $\widetilde{L} = \max\left\{4L\sqrt{\frac{\omega_{\rm dwn}}{\alpha_{\rm dwn}}\left(\frac{1}{\alpha_{\rm dwn}}+\frac{\omega_{\rm up}}{N}\right)},\ 4L\left(1+\frac{\omega_{\rm up}}{N}\right)\right\}$, taking for all k in $\mathbb{N}$, $\gamma_k = \frac{2}{\mu(k+1)+\widetilde{L}}$, for all k in N^*, $\lambda_k = \frac{1}{\gamma_{k-1}}$ and denoting $\bar{w}_K = \frac{\sum_{k=1}^{K}\lambda_k w_{k-1}}{\sum_{k=1}^{K}\lambda_k}$, then for any K in $\mathbb{N}^*$, we have:

$$\mathbb{E}\left[F(\bar{w}_K)-F(w_*)\right] \leq \frac{\mu+2\widetilde{L}}{4\mu K^2}\left\|w_0-w_*\right\|^2+\left(1+\frac{64L\omega_{\rm dwn}^2}{\mu K}\ln\left(\mu K+\widetilde{L}\right)\right)\cdot\frac{4\sigma^2(1+\omega_{\rm up})}{\mu K N b}\,,$$

and the demonstration is completed.

□

E.4 Non-convex case (extra theorem)

In this section, we detail the convergence guarantee given for MCM in the non-convex case. In this scenario, the theorem will hold on the average of gradients after K in $\mathbb{N}^*$ iterations. The structure of the proof is different from the one used for Ghost and MCM in convex and strongly-convex case. Instead, the demonstration starts from the equation resulting from smoothness and use the polarization identity to handle the inner product of gradients taken at two different points.

Theorem S11 (Convergence of MCM in the non-convex case)**.** *Under Assumptions 1, 2 and 4 (non-convex case), for a learning rate* $\alpha_{\rm dwn} = \frac{1}{8\omega_{\rm dwn}}$*, for any step size* γ *s.t.*

$$\gamma L \leq \min\left\{\frac{1}{8\omega_{\rm dwn}},\ \frac{1}{2\left(1+\frac{\omega_{\rm up}}{N}\right)},\ \frac{1}{8\sqrt{\omega_{\rm dwn}^2\left(8\omega_{\rm dwn}+\frac{\omega_{\rm up}}{N}\right)}}\right\}\,,$$

after running K *in* $\mathbb{N}^*$ *iterations, we have:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(w_{k-1})\right\|^2\right] \leq \frac{2\left(F(w_0)-F(w_*)\right)}{\gamma K}+\frac{\gamma L\sigma^2\Phi^{\rm non-cvx}(\gamma)}{Nb}\,,$$

with $\Phi^{\rm non-cvx}(\gamma) := (1+\omega_{\rm up})\left(1+32\gamma L\omega_{\rm dwn}^2\right)$*. Thus, for* K *in* $\mathbb{N}^*$ *large enough, taking* $\gamma = \sqrt{\frac{2Nb\left(F(w_0)-F(w_*)\right)}{\sigma^2 L(1+\omega_{\rm up})K}}$*:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(w_{k-1})\right\|^2\right] \leq 2\sqrt{\frac{2L\sigma^2(1+\omega_{\rm up})\left(F(w_0)-F(w_*)\right)}{NbK}}+O(K^{-1})\,.$$

Proof. Let k in $\mathbb{N}^*$, then smoothness (see Assumption 2) implies:

$$\begin{aligned} & F(w_k) \leq F(w_{k-1}) + \langle\, \nabla F(w_{k-1}),\, w_k - w_{k-1} \,\rangle + \frac{L}{2} \left\| w_k - w_{k-1} \right\|^2 \\ \iff \quad & F(w_k) \leq F(w_{k-1}) - \gamma \,\langle\, \nabla F(w_{k-1}),\, \widetilde{g}_k \,\rangle + \frac{\gamma^2 L}{2} \left\| \widetilde{g}_k \right\|^2 . \end{aligned}$$

The inner product is not easy to handle because it implies two gradients computed at two different points: w_{k-1} and $\widehat{w}_{k-1}$. To turn around this difficulty, we use the polarization identity, and so we have:

$$\begin{aligned} -\mathbb{E}\left[\langle\, \nabla F(w_{k-1}),\, \widetilde{g}_k \,\rangle \mid w_{k-1}\right] &= -\,\langle\, \nabla F(w_{k-1}),\, \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right] \,\rangle \\ &= \frac{1}{2}\left(- \left\|\nabla F(w_{k-1})\right\|^2 - \mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \;\middle|\; w_{k-1} \right] \right. \\ &\qquad \left. + \mathbb{E}\left[\left\|\nabla F(w_{k-1}) - \nabla F(\widehat{w}_{k-1})\right\|^2 \;\middle|\; w_{k-1} \right] \right) \end{aligned}$$

where we used the Polarization identity (eq. (S6)), and next with smoothness:

$$\begin{aligned} -\mathbb{E}\left[\langle\, \nabla F(w_{k-1}),\, \widetilde{g}_k \,\rangle \mid w_{k-1}\right] \leq \frac{1}{2}\left(- \left\|\nabla F(w_{k-1})\right\|^2 - \mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \;\middle|\; w_{k-1} \right] \right. \\ \left. + L^2 \mathbb{E}\left[\left\| w_{k-1} - \widehat{w}_{k-1}\right\|^2 \;\middle|\; w_{k-1} \right] \right) , \end{aligned}$$

Combining with Lemma S3, we obtain:

$$\begin{aligned} F(w_k) \leq F(w_{k-1}) &- \frac{\gamma}{2} \left\|\nabla F(w_{k-1})\right\|^2 - \frac{\gamma}{2} \mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \;\middle|\; w_{k-1} \right] \\ &+ \frac{\gamma L^2}{2} \mathbb{E}\left[\left\| w_{k-1} - \widehat{w}_{k-1}\right\|^2 \;\middle|\; w_{k-1} \right] \\ &+ \frac{\gamma^2 L}{2} \left(\left(1 + \frac{\omega_{\mathrm{up}}}{N} \right) \left\|\nabla F(\widehat{w}_{k-1})\right\|^2 + \frac{\sigma^2 (1 + \omega_{\mathrm{up}})}{Nb} \right) . \end{aligned}$$

Taking the full expectation and re-ordering the terms gives:

$$\begin{aligned} \mathbb{E}\left[F(w_k)\right] \leq \mathbb{E}\left[F(w_{k-1})\right] &- \frac{\gamma}{2} \mathbb{E}\left[\left\|\nabla F(w_{k-1})\right\|^2 \right] - \frac{\gamma}{2} \left(1 - \gamma L \left(1 + \frac{\omega_{\mathrm{up}}}{N} \right) \right) \mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \right] \\ &+ \frac{\gamma L^2}{2} \mathbb{E}\left[\left\| w_{k-1} - \widehat{w}_{k-1}\right\|^2 \right] + \frac{\gamma^2 L}{2} \times \frac{\sigma^2 (1 + \omega_{\mathrm{up}})}{Nb} . \end{aligned}$$

Exactly like the convex case, we consider that $\gamma L(1 + \omega_{\mathrm{up}}/N) \leq 1/2$ and because $\mathbb{E}\left[\left\| w_{k-1} - \widehat{w}_{k-1}\right\|^2 \right] = \mathbb{E}\left[\mathbb{E}\left[\left\| w_{k-1} - \widehat{w}_{k-1}\right\|^2 \;\middle|\; \widehat{w}_{k-2} \right] \right]$ we can use Assumption 1:

$$\begin{aligned} \mathbb{E}\left[F(w_k)\right] \leq \mathbb{E}\left[F(w_{k-1})\right] &- \frac{\gamma}{2} \mathbb{E}\left[\left\|\nabla F(w_{k-1})\right\|^2 \right] - \frac{\gamma}{4} \mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \right] \\ &+ \frac{\omega_{\mathrm{dwn}} \gamma L^2}{2} \mathbb{E}\left[\Upsilon_k\right] + \frac{\gamma^2 L}{2} \times \frac{\sigma^2 (1 + \omega_{\mathrm{up}})}{Nb} . \end{aligned} \tag{S22}$$

Next, Theorem 3 gives:

$$\mathbb{E}\left[\Upsilon_k\right] \leq \left(1 - \frac{\alpha_{\mathrm{dwn}}}{2} \right) \mathbb{E}\left[\Upsilon_{k-1}\right] + 2\gamma^2 \left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{\omega_{\mathrm{up}}}{N} \right) \mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2 \right] + \frac{2\gamma^2 \sigma^2 (1 + \omega_{\mathrm{up}})}{Nb} .$$

We iterate over k and compute the resulting geometric sum, it gives:

$$\begin{aligned} \mathbb{E}\left[\Upsilon_k\right] \leq \left(1 - \frac{\alpha_{\mathrm{dwn}}}{2} \right)^k \left\|\Upsilon_0\right\|^2 + 2\gamma^2 \left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{\omega_{\mathrm{up}}}{N} \right) \sum_{t=1}^{k} \left(1 - \frac{\alpha}{2} \right)^{k-t} \mathbb{E}\left[\left\|\nabla F(\widehat{w}_{t-1})\right\|^2 \right] \\ + \frac{4\gamma^2 \sigma^2 (1 + \omega_{\mathrm{up}})}{\alpha_{\mathrm{dwn}} Nb} , \end{aligned}$$

where we considered for the last term of the above equation that $\sum_{t=1}^{k}\left(1-\frac{\alpha_{\text{dwn}}}{2}\right)^{k} \leq \frac{2}{\alpha_{\text{dwn}}}$. This is equivalent to:

$$\mathbb{E}\left[\Upsilon_k\right] \leq 2\gamma^2\left(\frac{1}{\alpha_{\text{dwn}}}+\frac{\omega_{\text{up}}}{N}\right)\sum_{t=1}^{k}\left(1-\frac{\alpha_{\text{dwn}}}{2}\right)^{k-t}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{t-1})\right\|^2\right]+\frac{4\gamma^2\sigma^2(1+\omega_{\text{up}})}{\alpha_{\text{dwn}}Nb}\,.$$

We apply this last result to eq. (S22):

$$\begin{aligned}
\frac{\gamma}{2}\mathbb{E}\left[\left\|\nabla F(w_{k-1})\right\|^2\right] &\leq \mathbb{E}\left[F(w_{k-1})-F(w_k)\right]-\frac{\gamma}{4}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right] \\
&\quad+\frac{\gamma L^2}{2}\left(\frac{4\omega_{\text{dwn}}\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb\alpha_{\text{dwn}}}\right. \\
&\qquad\left.+2\omega_{\text{dwn}}\gamma^2\left(\frac{1}{\alpha_{\text{dwn}}}+\frac{\omega_{\text{up}}}{N}\right)\sum_{t=1}^{k}\left(1-\frac{\alpha_{\text{dwn}}}{2}\right)^{k-t}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{t-1})\right\|^2\right]\right) \\
&\quad+\frac{\gamma^2 L}{2}\times\frac{\sigma^2(1+\omega_{\text{up}})}{Nb} \\
&\leq \mathbb{E}\left[F(w_{k-1})-F(w_k)\right]-\frac{\gamma}{4}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right] \\
&\quad+\gamma^3L^2\omega_{\text{dwn}}\left(\frac{1}{\alpha_{\text{dwn}}}+\frac{\omega_{\text{up}}}{N}\right)\sum_{t=1}^{k}\left(1-\frac{\alpha_{\text{dwn}}}{2}\right)^{k-t}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{t-1})\right\|^2\right] \\
&\quad+\frac{\gamma^2\sigma^2L(1+\omega_{\text{up}})}{2Nb}\left(1+\frac{4\gamma L\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\right).
\end{aligned}$$

Summing this equation, for k in range 1 to K:

$$\begin{aligned}
\frac{\gamma}{2}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(w_{k-1})\right\|^2\right] &\leq \mathbb{E}\left[F(w_0)-F(w_k)\right]-\frac{\gamma}{4}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right] \\
&\quad+\gamma^3L^2\omega_{\text{dwn}}\left(\frac{1}{\alpha_{\text{dwn}}}+\frac{\omega_{\text{up}}}{N}\right)\sum_{k=1}^{K}\sum_{t=1}^{k}\left(1-\frac{\alpha_{\text{dwn}}}{2}\right)^{k-t}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{t-1})\right\|^2\right] \\
&\quad+\frac{\gamma^2\sigma^2L(1+\omega_{\text{up}})}{2Nb}\left(1+\frac{4\gamma L\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\right)K\,.
\end{aligned}$$

We need to invert the double-sum and we obtain:

$$\begin{aligned}
\frac{\gamma}{2}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(w_{k-1})\right\|^2\right] &\leq \gamma F(w_0)-F(w_k)-\frac{\gamma}{4}\sum_{i=1}^{K}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right] \\
&\quad+\frac{2}{\alpha_{\text{dwn}}}\times\gamma^3L^2\omega_{\text{dwn}}\left(\frac{1}{\alpha_{\text{dwn}}}+\frac{\omega_{\text{up}}}{N}\right)\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right] \\
&\quad+\frac{\gamma^2\sigma^2L(1+\omega_{\text{up}})}{2Nb}\left(1+\frac{4\gamma L\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\right)K \\
&\leq \mathbb{E}\left[F(w_0)-F(w_k)\right] \\
&\quad+\left(2\gamma^3L^2\frac{\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\left(\frac{1}{\alpha_{\text{dwn}}}+\frac{\omega_{\text{up}}}{N}\right)-\frac{\gamma}{4}\right)\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1})\right\|^2\right] \\
&\quad+\frac{\gamma^2\sigma^2L(1+\omega_{\text{up}})}{2Nb}\left(1+\frac{4\gamma L\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\right)K\,.
\end{aligned}$$

Now we consider that $2\gamma^3 L^2 \frac{\omega_{\rm dwn}}{\alpha_{\rm dwn}} \left(\frac{1}{\alpha_{\rm dwn}} + \frac{\omega_{\rm up}}{N}\right) \leq \gamma/4$, and because for all k in $\mathbb{N}$, $F(w_0) - F(w_k) \leq F(w_0) - F(w_*)$:

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\left\|\nabla F(w_{k-1})\right\|^2\right] \leq \frac{2\left(F(w_0) - F(w_*)\right)}{\gamma K} + \frac{\gamma\sigma^2 L(1+\omega_{\rm up})}{Nb}\left(1 + \frac{4\gamma L\omega_{\rm dwn}}{\alpha_{\rm dwn}}\right).$$

Finally, for any K in $\mathbb{N}^*$, such that $\gamma L \leq \min\left\{\frac{1}{8\omega_{\rm dwn}}, \frac{1}{2\left(1+\frac{\omega_{\rm up}}{N}\right)}, \frac{1}{2\sqrt{2\frac{\omega_{\rm dwn}}{\alpha_{\rm dwn}}\left(\frac{1}{\alpha_{\rm dwn}} + \frac{\omega_{\rm up}}{N}\right)}}\right\}$ and $\alpha_{\rm dwn} \leq \frac{1}{8\omega_{\rm dwn}}$, we have:

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\left\|\nabla F(w_{k-1})\right\|^2\right] \leq \frac{2\left(F(w_0) - F(w_*)\right)}{\gamma K} + \frac{\gamma L\sigma^2 \Phi^{\rm non-cvx}(\gamma)}{Nb},$$

denoting $\Phi^{\rm non-cvx}(\gamma) := (1+\omega_{\rm up})\left(1 + \frac{4\gamma L\omega_{\rm dwn}}{\alpha_{\rm dwn}}\right)$.

Thus, for K in $\mathbb{N}^*$ large enough, taking $\gamma = \sqrt{\frac{2Nb\left(F(w_0) - F(w_*)\right)}{\sigma^2 L(1+\omega_{\rm up})K}}$ and $\alpha_{\rm dwn} = 1/(8\omega_{\rm dwn})$:

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\left\|\nabla F(w_{k-1})\right\|^2\right] \leq 2\sqrt{\frac{2L\sigma^2(1+\omega_{\rm up})\left(F(w_0) - F(w_*)\right)}{NbK}} + O(K^{-1}).$$

□

E.5 Proof for `Rand-MCM` (Theorem 4)

The proof for `Rand-MCM` is almost identical to the `MCM`-scenario. It only requires to modify some notations because each device i in $[\![1, N]\!]$ holds a unique model $\widehat{w}_{k-1}^i$.

For k in $\mathbb{N}$:

1. $\widetilde{g}_k$ is now defined as $\widetilde{g}_k = \frac{1}{N}\sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}^i)$,
2. for all i in $[\![1, N]\!]$, $\widehat{g}_k^i(\widehat{w}_{k-1})$ and $\nabla F(\widehat{w}_{k-1})$ must be replaced by $\widehat{g}_k^i(\widehat{w}_{k-1}^i)$ and $\nabla F(\widehat{w}_{k-1}^i)$,
3. instead of having a unique memory H_k, there is N memories $(H_k^i)_{i=1}^N$ that keep track of the updates done on each worker,
4. furthermore the notation $w_{k-1} - H_{k-2}$ is no more correct as we have N different memories. Thus, it must be replaced by $\frac{1}{N}\sum_{i=1}^N w_{k-1} - H_{k-2}^i$.

F Proofs in the quadratic case for `MCM` and `Rand-MCM`

In this section, for ease of notation we denote for k in $\mathbb{N}^*$, $\widetilde{g}_k = \frac{1}{N}\sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}^i)$.

`MCM` has a unique memory H_k, and `Rand-MCM` has N different memories $(H_k^i)_{i=1}^N$. But for the sake of factorization, we will consider that both algorithm have N memories, thus we will always consider the quantity $\frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}^i\right\|^2$, while we should consider the quantity $\frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}\right\|^2$ for `MCM`. However this notation is correct considering that for `MCM`, for all i in $[\![1, N]\!]$, $H_k^i = H_k$. And it follows that we have $\frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}^i\right\|^2 = \frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}^i\right\|^2$.

Unlike the previous sections where the proofs for MCM and Rand-MCM do not require any distinction, here in the quadratic case, we will on the contrary stress on the difference between the two. The difference appears in Lemma S4 and comes from the way we handle the expectation of $\left\| \frac{1}{N} \sum_{i=1}^N \nabla F(\widehat{w}_{k-1}^i) - \nabla F(w_{k-1}) \right\|^2$ for k in $\mathbb{N}^*$. For this purpose we define a constant $\mathbf{C}$ such that $\mathbf{C} = 1$ in the MCM-case and $\mathbf{C} = N$ in the Rand-MCM-case.

The proofs for quadratic functions relies on the fact that for any k in $\mathbb{N}^*$, $\mathbb{E}\left[\nabla F(\widehat{w}_{k-1} \mid w_{k-1}\right] = \nabla F(w_{k-1})$.

Definition 2 (Quadratic function)**.** *A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is said to be quadratic if there exists a symmetric matrix A in $\mathcal{M}_{d,d}(\mathbb{R})$ such that for all x in $\mathbb{R}^d$: $f(x) - f(x_*) = \frac{1}{2}(x - x_*)^T A(x - x_*)$. And then its gradient is defined for all x in $\mathbb{R}^d$ as: $\nabla f(x) = A(x - x_*)$.*

F.1 Two other lemmas

In this section, we detail two lemmas required to prove the convergence of MCM and Rand-MCM in the case of quadratic functions.

The first lemma allows to factorize all the results obtained for both MCM and Rand-MCM algorithms. For k in $\mathbb{N}^*$ and i in $[\![1, N]\!]$, the difference between the MCM-case and the Rand-MCM-case results from the tigher control of $\left\| \sum_{i=1}^N \nabla F(\widehat{w}_{k-1}^i) - \nabla F(w_{k-1}) \right\|^2$.

Lemma S4. *We define $\mathbf{C}$ such that $\mathbf{C} = 1$ in the MCM-case and $\mathbf{C} = N$ in the Rand-MCM-case. Then for any k in $\mathbb{N}^*$, we have:*

$$\mathbb{E}\left[\left\| \frac{1}{N} \sum_{i=1}^N \nabla F(\widehat{w}_{k-1}^i) - \nabla F(w_{k-1}) \right\|^2 \,\middle|\, w_{k-1} \right] \leq \frac{L^2 \omega_{\mathrm{dwn}}}{\mathbf{C}} \frac{1}{N} \sum_{i=1}^N \left\| w_{k-1} - H_{k-2}^i \right\|^2 .$$

Proof. Let k in $\mathbb{N}^*$, we apply smoothness (see Assumption 2), and then we upper bound the variance of the quantization operator with Assumption 1. But we must distinguish MCM and Rand-MCM because in the first case we have $\widehat{w}_{k-1}^i$ equal to $\widehat{w}_{k-1}$ for all i in $[\![1, N]\!]$.

In the MCM-case:

$$\begin{aligned}
\mathbb{E}\left[\left\| \frac{1}{N} \sum_{i=1}^N \nabla F(\widehat{w}_{k-1}^i) - \nabla F(w_{k-1}) \right\|^2 \,\middle|\, w_{k-1} \right] &= \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) - F(w_{k-1}) \mid w_{k-1}\right] \\
&\leq L^2 \mathbb{E}\left[\left\| \widehat{w}_{k-1} - w_{k-1} \right\|^2 \,\middle|\, w_{k-1} \right] \\
&\leq L^2 \omega_{\mathrm{dwn}} \left\| \Omega_{k-1} \right\|^2 \\
&\leq L^2 \omega_{\mathrm{dwn}} \frac{1}{N} \sum_{i=1}^N \left\| w_{k-1} - H_{k-2}^i \right\|^2 ,
\end{aligned}$$

because we consider that $\left\| \Omega_{k-1} \right\|^2 = \left\| w_{k-1} - H_{k-2} \right\|^2 = \frac{1}{N} \sum_{i=1}^N \left\| w_{k-1} - H_{k-2}^i \right\|^2$.

In the Rand-MCM-case, by independence of the compressions on the downlink direction:

$$\begin{aligned}
\mathbb{E}\left[\left\| \frac{1}{N} \sum_{i=1}^N \nabla F(\widehat{w}_{k-1}^i) - \nabla F(w_{k-1}) \right\|^2 \,\middle|\, w_{k-1} \right] &= \frac{1}{N^2} \sum_{i=1}^N \mathbb{E}\left[\left\| \nabla F(\widehat{w}_{k-1}^i) - \nabla F(w_{k-1}) \right\|^2 \,\middle|\, w_{k-1} \right] \\
&\leq \frac{L^2}{N^2} \sum_{i=1}^N \left\| \widehat{w}_{k-1}^i - w_{k-1} \right\|^2 \\
&\leq \frac{L^2 \omega_{\mathrm{dwn}}}{N} \times \frac{1}{N} \sum_{i=1}^N \left\| w_{k-1} - H_{k-2}^i \right\|^2 \\
&\leq \frac{L^2 \omega_{\mathrm{dwn}}}{N} \frac{1}{N} \sum_{i=1}^N \left\| w_{k-1} - H_{k-2}^i \right\|^2 .
\end{aligned}$$

We factorize the two results and define $\mathbf{C}$ such that $\mathbf{C} = 1$ in the `MCM`-case and $\mathbf{C} = N$ in the `Rand-MCM`-case, and the result follows.

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F(\widehat{w}_{k-1}^{i}) - \nabla F(w_{k-1})\right\|^2 \,\middle|\, w_{k-1}\right] \leq \frac{L^2\omega_{\mathrm{dwn}}}{\mathbf{C}}\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1} - H_{k-2}^{i}\right\|^2 .$$

□

The next lemma replaces Lemma S3 in the context of randomization and quadratic functions. Note that the conditioning in Lemma S3 is w.r.t. to $\widehat{w}_{k-1}$ while here we take the expectation w.r.t. w_{k-1}. This is because we remove $\widehat{w}_{k-1}$ from the gradient and give a result which depends of $\|\nabla F(w_{k-1})\|^2$ instead of $\|\nabla F(\widehat{w}_{k-1})\|^2$. This is made possible by the fact that for all k in $\mathbb{N}$, for quadratic functions, we have $\mathbb{E}\left[\nabla F(\widehat{w}_{k-1})\right] = \nabla F(w_{k-1})$.

Lemma S5 (Squared-norm of stochastic gradients)**.** *For any k in $\mathbb{N}^*$, the squared-norm of gradients can be bounded a.s.:*

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\widehat{g}_k^i(\widehat{w}_{k-1}^{i}) - \nabla F(\widehat{w}_{k-1}^{i})\right\|^2 \,\middle|\, w_{k-1}\right] \leq \frac{\omega_{\mathrm{up}}}{N}\left\|\nabla F(w_{k-1})\right\|^2 + \frac{\sigma^2(1+\omega_{\mathrm{up}})}{Nb} + \frac{\omega_{\mathrm{up}}\omega_{\mathrm{dwn}}L^2}{N}\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1} - H_{k-2}^{i}\right\|^2 , \tag{S23}$$

$$\mathbb{E}\left[\left\|\widetilde{g}_k\right\|^2 \,\middle|\, w_{k-1}\right] \leq \left(1 + \frac{\omega_{\mathrm{up}}}{N}\right)\left\|\nabla F(w_{k-1})\right\|^2 + \frac{\sigma^2(1+\omega_{\mathrm{up}})}{Nb} + L^2\omega_{\mathrm{dwn}}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\mathrm{up}}}{N}\right)\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1} - H_{k-2}^{i}\right\|^2 . \tag{S24}$$

The demonstration will be in two stages. We first show eq. (S23), and in a second time, we show eq. (S24).

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

First part (eq. (S23)). We can decompose the squared-norm in two terms:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{g}_k^i(\widehat{w}_{k-1}^{i}) - \nabla F(\widehat{w}_{k-1}^{i})\right)\right\|^2 \,\middle|\, w_{k-1}\right] \\
&\quad = \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{g}_k^i(\widehat{w}_{k-1}^{i}) - g_k^i(\widehat{w}_{k-1}^{i})\right)\right\|^2 \,\middle|\, w_{k-1}\right] \\
&\quad + \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\left(g_k^i(\widehat{w}_{k-1}^{i}) - \nabla F(\widehat{w}_{k-1}^{i})\right)\right\|^2 \,\middle|\, w_{k-1}\right] ,
\end{aligned}$$

the first term is bounded by Assumption 1 and the last term by Assumption 4:

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{g}_k^i(\widehat{w}_{k-1}^i)-\nabla F(\widehat{w}_{k-1}^i)\right)\right\|^2 \;\middle|\; w_{k-1}\right] &\leq \frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_k^i(\widehat{w}_{k-1}^i)\right\|^2 \;\middle|\; w_{k-1}\right]+\frac{\sigma^2}{Nb}\\
&\leq \frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|g_k^i(\widehat{w}_{k-1}^i)-\nabla F(\widehat{w}_{k-1}^i)\right\|^2 \;\middle|\; w_{k-1}\right]\\
&\quad+\frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1}^i)\right\|^2 \;\middle|\; w_{k-1}\right]+\frac{\sigma^2}{Nb}\,.
\end{aligned}$$

And again applying Assumption 4 on $\mathbb{E}\left[\left\|g_k^i(\widehat{w}_{k-1}^i)-\nabla F(\widehat{w}_{k-1}^i)\right\|^2 \;\middle|\; w_{k-1}\right]$ for i in $\{1,\cdots N\}$:

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{g}_k^i(\widehat{w}_{k-1}^i)-\nabla F(\widehat{w}_{k-1}^i)\right)\right\|^2 \;\middle|\; w_{k-1}\right] &= \frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1}^i)\right\|^2 \;\middle|\; w_{k-1}\right]\\
&\quad+\frac{\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\,.
\end{aligned}$$

Now, we have:

$$\begin{aligned}
\frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1}^i)\right\|^2 \;\middle|\; w_{k-1}\right] &= \frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1}^i)-\nabla F(w_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right]\\
&\quad+\frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla F(w_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right]\,,
\end{aligned}$$

using smoothness (Assumption 2) gives:

$$\frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla F(\widehat{w}_{k-1}^i)\right\|^2 \;\middle|\; w_{k-1}\right] = \frac{\omega_{\mathrm{up}}\omega_{\mathrm{dwn}}L^2}{N}\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1}-H_{k-2}^i\right\|^2+\frac{\omega_{\mathrm{up}}}{N}\left\|\nabla F(w_{k-1})\right\|^2\,,$$

and putting everythings together allows to conclude for eq. (S23).

Second part (eq. (S24)). We start by introducing $\|\nabla F(w_{k-1})\|^2$:

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\widehat{g}_k^i(\widehat{w}_{k-1}^i)\right\|^2 \;\middle|\; w_{k-1}\right] &= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\widehat{g}_k^i(\widehat{w}_{k-1}^i)-\nabla F(w_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right]\\
&\quad+\left\|\nabla F(w_{k-1})\right\|^2\\
&= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\widehat{g}_k^i(\widehat{w}_{k-1}^i)-\nabla F(\widehat{w}_{k-1}^i)\right\|^2 \;\middle|\; w_{k-1}\right]\\
&\quad+\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F(\widehat{w}_{k-1}^i)-\nabla F(w_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right]\\
&\quad+\left\|\nabla F(w_{k-1})\right\|^2\,.
\end{aligned}$$

The second term of the previous line is controlled by Lemma S4 which distinguish the `MCM` and `Rand-MCM`-cases by defining a constant $\mathbf{C}$ such that $\mathbf{C}=1$ for `MCM` and $\mathbf{C}=N$ for `Rand-MCM`:

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F(\widehat{w}_{k-1}^i)-\nabla F(w_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right] \leq \frac{L^2\omega_{\mathrm{dwn}}}{\mathbf{C}}\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1}-H_{k-2}^i\right\|^2\,.$$

Thus, we have:

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\widehat{g}_k^i(\widehat{w}_{k-1}^i)\right\|^2 \,\middle|\, w_{k-1}\right] = \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\widehat{g}_k^i(\widehat{w}_{k-1}^i) - \nabla F(\widehat{w}_{k-1}^i)\right\|^2 \,\middle|\, w_{k-1}\right] + \frac{\omega_{\mathrm{dwn}}L^2}{\mathbf{C}}\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1} - H_{k-2}^i\right\|^2 + \|\nabla F(w_{k-1})\|^2 ,$$

and eq. (S23) allows to conclude. □

F.2 Control of the Variance of the local model for quadratic function (both `MCM` and `Rand-MCM`)

The next theorem replaces the Theorem 3 in the case of quadratic functions. The results are almost identical except that in these settings we control the variance using non-degraded points $(w_t)_{t\in\mathbb{N}}$. This is necessary because, for quadratic functions, the analysis is slightly different. Previously, we upper-bounded the inner product in the decomposition (eq. (S12)) by a "strong contraction" that was allowing to subtract $\|\nabla F(\widehat{w}_{k-1})\|^2$ and an extra residual term. Here we instead directly get a smaller contraction proportional to $\|\nabla F(w_{k-1})\|^2$ (but without any residual!). Indeed for all k in $\mathbb{N}$, we have $\mathbb{E}\left[\nabla F(\widehat{w}_{k-1})\right] = \nabla F(w_{k-1})$. This difference will appear in Appendix F.3.

As a consequence, we need to also control the variance of the local iterates that will appear when expanding the expected squared gradient $\mathbb{E}\|\tilde{g}_k\|^2$ by an affine function of the squared norms of the gradients **at the non perturbed points**. This is what Theorem S12 provides.

Theorem S12. *Consider the `MCM` update as in eq. (2) or the `Rand-MCM` update as described in Subsection 2.2. Under Assumptions 1 to 4 with $\mu = 0$, if $\gamma \leq \dfrac{1}{8L\omega_{\mathrm{dwn}}\sqrt{(1/\mathbf{C} + \omega_{\mathrm{up}}/N)}}$ and $\alpha_{\mathrm{dwn}} \leq 1/(8\omega_{\mathrm{dwn}})$, then for all k in $\mathbb{N}$:*

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|w_k - H_{k-1}^i\right\|^2 \,\middle|\, w_{k-1}\right] \leq 2\gamma^2\left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{\omega_{\mathrm{up}}}{N}\right)\sum_{t=1}^{k}(1 - \frac{\alpha_{\mathrm{dwn}}}{2})^{k-t}\mathbb{E}\left[\|\nabla F(w_{t-1})\|^2 \,\middle|\, w_{t-1}\right] + \frac{4\gamma^2\sigma^2(1+\omega_{\mathrm{up}})}{\alpha_{\mathrm{dwn}}Nb} .$$

Proof. Let k in $\mathbb{N}^*$ and i in $\{1, \dots N\}$, from Theorem S8 we have:

$$\mathbb{E}\left[\left\|w_k - H_{k-1}^i\right\|^2 \,\middle|\, w_{k-1}\right] = \mathrm{Var} + \mathrm{Bias}^2 = 2\gamma^2\mathrm{Var}_1 + 2\alpha_{\mathrm{dwn}}^2\mathrm{Var}_2 + \mathrm{Bias}^2 ,$$

with

$$\begin{cases} \mathrm{Var}_1 & = \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\widehat{g}_k^i(\widehat{w}_{k-1}^i) + \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}^i) \mid w_{k-1}\right]\right\|^2 \,\middle|\, w_{k-1}\right] \\ \mathrm{Var}_2 & = \omega_{\mathrm{dwn}}\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1} - H_{k-2}^i\right\|^2 \\ \mathrm{Bias}^2 & = \left\|\mathbb{E}\left[w_k - H_{k-1} \mid w_{k-1}\right]\right\|^2 . \end{cases}$$

Recall that in the case of quadratic functions, we have for all i in $[\![1, N]\!]$: $\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}^i) \mid w_{k-1}\right] = \nabla F(w_{k-1})$. And so for the first term of variance we can decompose as following:

$$\begin{aligned}
\text{Var}_1 &= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}^i) - \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}^i) \mid w_{k-1}\right]\right\|^2 \;\middle|\; w_{k-1}\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}^i) - \nabla F(w_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}^i) - \nabla F(\widehat{w}_{k-1}^i)\right\|^2 \;\middle|\; w_{k-1}\right] \\
&\quad + \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \nabla F(\widehat{w}_{k-1}^i) - \nabla F(w_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right].
\end{aligned}$$

The first part is handled by eq. (S23) of Lemma S5:

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \widehat{g}_k^i(\widehat{w}_{k-1}^i) - \nabla F(\widehat{w}_{k-1}^i)\right\|^2 \;\middle|\; w_{k-1}\right] &= \frac{\omega_{\text{up}}\omega_{\text{dwn}}L^2}{N}\frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}^i\right\|^2 \\
&\quad + \frac{\omega_{\text{up}}}{N}\left\|\nabla F(w_{k-1})\right\|^2 \\
&\quad + \frac{\sigma^2(1+\omega_{\text{up}})}{Nb},
\end{aligned}$$

and the second part is tackled by Lemma S4 where is defined a constant $\mathbf{C}$ such that $\mathbf{C} = 1$ in the `MCM`-case, and $\mathbf{C} = N$ in the `Rand-MCM`-case: $\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \nabla F(\widehat{w}_{k-1}^i) - \nabla F(w_{k-1})\right\|^2 \;\middle|\; w_{k-1}\right] \leq \frac{L^2\omega_{\text{dwn}}}{\mathbf{C}}\frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}^i\right\|^2$.

Finally, given that $\text{Var} = 2\gamma^2\text{Var}_1 + 2\alpha_{\text{dwn}}^2\text{Var}_2$ we have:

$$\begin{aligned}
\text{Var} &\leq 2\gamma^2L^2\omega_{\text{dwn}}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\text{up}}}{N}\right)\frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}^i\right\|^2 + 2\alpha_{\text{dwn}}^2\omega_{\text{dwn}}\left\|w_{k-1} - H_{k-2}^i\right\|^2 \\
&\quad + \frac{2\gamma^2\omega_{\text{up}}}{N}\left\|\nabla F(w_{k-1})\right\|^2 + \frac{2\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb}.
\end{aligned}$$

Now we focus on the squared bias Bias^2 exactly like in Theorem S8 and we obtain:

$$\text{Bias}^2 \leq (1-\alpha_{\text{dwn}})\left\|w_{k-1} - H_{k-2}^i\right\|^2 + \gamma^2(1 + \frac{1}{\alpha_{\text{dwn}}})\left\|\nabla F(w_{k-1})\right\|^2.$$

At the end:

$$\begin{aligned}
\mathbb{E}\left[\left\|w_k - H_{k-1}^i\right\|^2 \;\middle|\; w_{k-1}\right] &\leq 2\gamma^2L^2\omega_{\text{dwn}}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\text{up}}}{N}\right)\frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}^i\right\|^2 \\
&\quad + \gamma^2(1 + \frac{1}{\alpha_{\text{dwn}}} + \frac{2\omega_{\text{up}}}{N})\left\|\nabla F(w_{k-1})\right\|^2 \\
&\quad + \left((1-\alpha_{\text{dwn}}) + 2\alpha_{\text{dwn}}^2\omega_{\text{dwn}}\right)\left\|w_{k-1} - H_{k-2}^i\right\|^2 \\
&\quad + \frac{2\gamma^2\sigma^2(1+\omega_{\text{up}})}{Nb}.
\end{aligned}$$

Summing this last equation over the N devices gives:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|w_k - H_{k-1}^i\right\|^2 \,\middle|\, w_{k-1}\right]$$
$$\leq \left(1-\alpha_{\mathrm{dwn}}+2\alpha_{\mathrm{dwn}}^2\omega_{\mathrm{dwn}}+\gamma^2L^2\omega_{\mathrm{dwn}}\left(\frac{1}{\mathbf{C}}+\frac{\omega_{\mathrm{up}}}{N}\right)\right)\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1}-H_{k-2}^i\right\|^2$$
$$+\gamma^2(1+\frac{1}{\alpha_{\mathrm{dwn}}}+\frac{2\omega_{\mathrm{up}}}{N})\left\|\nabla F(w_{k-1})\right\|^2$$
$$+\frac{2\gamma^2\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\,.$$

Exactly like in Theorem S8, we need and by taking $\alpha_{\mathrm{dwn}}=1/(8\omega_{\mathrm{dwn}})$:

$$\begin{cases} 2\alpha_{\mathrm{dwn}}^2\omega_{\mathrm{dwn}} \leq \frac{1}{4}\alpha_{\mathrm{dwn}} \iff \alpha_{\mathrm{dwn}} \leq \frac{1}{8\omega_{\mathrm{dwn}}}\,, \\ 2\gamma^2L^2\omega_{\mathrm{dwn}}\left(\frac{1}{\mathbf{C}}+\frac{\omega_{\mathrm{up}}}{N}\right) \leq \frac{1}{4}\alpha_{\mathrm{dwn}} = \frac{1}{32\omega_{\mathrm{dwn}}} \iff \gamma \leq \frac{1}{8L\omega_{\mathrm{dwn}}\sqrt{(1/\mathbf{C}+\omega_{\mathrm{up}}/N)}}\,, \\ 1+\frac{1}{\alpha_{\mathrm{dwn}}} \leq \frac{2}{\alpha_{\mathrm{dwn}}} \text{ which is not restrictive.} \end{cases}$$

Thus, we can write:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|w_k - H_{k-1}^i\right\|^2 \,\middle|\, w_{k-1}\right] \leq \left(1-\frac{\alpha_{\mathrm{dwn}}}{2}\right)\frac{1}{N}\sum_{i=1}^{N}\left\|w_{k-1}-H_{k-2}^i\right\|^2$$
$$+2\gamma^2(\frac{1}{\alpha_{\mathrm{dwn}}}+\frac{\omega_{\mathrm{up}}}{N})\left\|\nabla F(w_{k-1})\right\|^2$$
$$+\frac{2\gamma^2\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\,.$$

Finally, we take the full expectation without any conditioning, we iterate over k and compute the geometric sums:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|w_k - H_{k-1}^i\right\|^2\right] \leq (1-\frac{\alpha_{\mathrm{dwn}}}{2})^k\left\|w_0-H_{-1}\right\|^2+\frac{4\gamma^2\sigma^2(1+\omega_{\mathrm{up}})}{\alpha_{\mathrm{dwn}}Nb}$$
$$+2\gamma^2(\frac{1}{\alpha_{\mathrm{dwn}}}+\frac{\omega_{\mathrm{up}}}{N})\sum_{t=1}^{k}(1-\frac{\alpha_{\mathrm{dwn}}}{2})^{k-t}\mathbb{E}\left[\left\|\nabla F(w_{t-1})\right\|^2\right]\,.$$

and the result follows.

□

F.3 Proof for quadratic function (Theorem 5)

Theorem S13. *Under Assumptions 1 to 4 with $\mu=0$, if the function is quadratic, for $\gamma=1/(L\sqrt{K})$ and a given learning rate $\alpha_{\mathrm{dwn}}=1/(8\omega_{\mathrm{dwn}})$, after running K iterations:*

$$\mathbb{E}\left[F(\bar{w}_K)-F_*\right] \leq \frac{\left\|w_0-w_*\right\|^2 L}{\sqrt{K}}+\frac{\sigma^2\Phi(\gamma)}{NbL\sqrt{K}}\,.$$

with $\Phi=(1+\omega_{\mathrm{up}})\left(1+32\frac{\omega_{\mathrm{dwn}}^2}{\sqrt{K}}\times\frac{1}{\sqrt{K}}\left(\frac{1}{\mathbf{C}}+\frac{\omega_{\mathrm{up}}}{N}\right)\right)$ and $\mathbf{C}=N$ for `Rand-MCM`, and 1 for `MCM`.

The structure of the proof is different from the one used in Appendices D and E.

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

$$\left\|w_k-w_*\right\|^2 \leq \left\|w_{k-1}-w_*\right\|^2-2\gamma\left\langle\,\widetilde{g}_k,\;w_{k-1}-w_*\,\right\rangle+\gamma^2\left\|\widetilde{g}_k\right\|^2\,.$$

Because F is quadratic, we have $\mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right] = \nabla F(w_{k-1})$, thus taking expectation gives:

$$\mathbb{E}\left[\|w_k - w_*\|^2 \;\middle|\; w_{k-1}\right] \leq \|w_{k-1} - w_*\|^2 - 2\gamma \left\langle \nabla F(w_{k-1}),\, w_{k-1} - w_* \right\rangle + \gamma^2 \mathbb{E}\left[\|\widetilde{g}_k\|^2 \;\middle|\; w_{k-1}\right].$$

We can directly apply convexity with eq. (S8) from Proposition S1:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2 \;\middle|\; w_{k-1}\right] \leq \|w_{k-1} - w_*\|^2 - \gamma\left(F(w_{k-1}) - F(w_*) + \frac{1}{L}\|\nabla F(w_{k-1})\|^2\right) \\ + \gamma^2 \mathbb{E}\left[\|\widetilde{g}_k\|^2 \;\middle|\; w_{k-1}\right].\end{aligned}$$

Now, with eq. (S24) of Lemma S5:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2 \;\middle|\; w_{k-1}\right] \leq \|w_{k-1} - w_*\|^2 - \gamma(F(w_{k-1}) - F(w_*)) - \frac{\gamma}{L}\|\nabla F(w_{k-1})\|^2 \\ + \gamma^2\left(\left(1 + \frac{\omega_{\mathrm{up}}}{N}\right)\|\nabla F(w_{k-1})\|^2 \right. \\ + L^2\omega_{\mathrm{dwn}}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\mathrm{up}}}{N}\right)\frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}^i\right\|^2 \\ \left. + \frac{\sigma^2(1+\omega_{\mathrm{up}})}{Nb}\right),\end{aligned}$$

which gives:

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2 \;\middle|\; w_{k-1}\right] \leq \|w_{k-1} - w_*\|^2 - \gamma(F(w_{k-1}) - F(w_*)) - \frac{\gamma}{L}\|\nabla F(w_{k-1})\|^2 \\ + \gamma^2\left(1 + \frac{\omega_{\mathrm{up}}}{N}\right)\|\nabla F(w_{k-1})\|^2 \\ + \gamma^2 L^2\omega_{\mathrm{dwn}}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\mathrm{up}}}{N}\right)\frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}^i\right\|^2 \\ + \frac{\sigma^2\gamma^2(1+\omega_{\mathrm{up}})}{Nb}.\end{aligned}$$

Taking full expectation, and because for all i in $\{1, \cdots, N\}$, $\mathbb{E}\left[\left\|w_{k-1} - H_{k-2}^i\right\|^2\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|w_{k-1} - H_{k-2}^i\right\|^2 \;\middle|\; \widehat{w}_{k-2}\right]\right]$, we can use the inequality controlling $\frac{1}{N}\sum_{i=1}^N \left\|w_{k-1} - H_{k-2}^i\right\|^2$ (see Theorem S12):

$$\begin{aligned}\mathbb{E}\left[\|w_k - w_*\|^2\right] \leq \mathbb{E}\left[\|w_{k-1} - w_*\|^2\right] - \gamma\mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] \\ - \frac{\gamma}{L}\left(1 - \gamma L\left(1 + \frac{\omega_{\mathrm{up}}}{N}\right)\right)\mathbb{E}\left[\|\nabla F(w_{k-1})\|^2\right] \\ + \gamma^2 L^2\omega_{\mathrm{dwn}}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\mathrm{up}}}{N}\right) \times 2\gamma^2\left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{\omega_{\mathrm{up}}}{N}\right)\sum_{t=1}^k (1 - \frac{\alpha_{\mathrm{dwn}}}{2})^{k-t}\mathbb{E}\left[\|\nabla F(w_{t-1})\|^2\right] \\ + \frac{\sigma^2\gamma^2(1+\omega_{\mathrm{up}})}{Nb} + \gamma^2 L^2\omega_{\mathrm{dwn}}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\mathrm{up}}}{N}\right) \times \frac{4\sigma^2\gamma^2(1+\omega_{\mathrm{up}})}{\alpha_{\mathrm{dwn}}Nb}.\end{aligned}$$

Next, we consider - as in previous proofs - that $\gamma L(1+\omega_{\rm up}/N) \leq 1/2$, and thus $\frac{\gamma}{L}\left(1-\gamma L\left(1+\frac{\omega_{\rm up}}{N}\right)\right) \geq \frac{\gamma}{2}$. Next we carry out the "top-down recurrence":

$$\begin{aligned}
\mathbb{E}\left[\|w_k - w_*\|^2\right] &\leq \|w_0 - w_*\|^2 - \gamma \sum_{j=1}^{k} \mathbb{E}\left[F(w_{k-j}) - F(w_*)\right] \\
&\quad - \frac{\gamma}{2L} \sum_{j=1}^{k} \mathbb{E}\left[\|\nabla F(w_{k-j-1})\|^2\right] \\
&\quad + \sum_{j=1}^{k} 2\gamma^4 L^2 \omega_{\rm dwn} \left(\frac{1}{\mathbf{C}} + \frac{\omega_{\rm up}}{N}\right)\left(\frac{1}{\alpha_{\rm dwn}} + \frac{\omega_{\rm up}}{N}\right) \sum_{t=1}^{k-j} \left(1 - \frac{\alpha_{\rm dwn}}{2}\right)^{k-j-t} \mathbb{E}\left[\|\nabla F(w_{t-1})\|^2\right] \\
&\quad + \sum_{j=1}^{k} \frac{\gamma^2\sigma^2(1+\omega_{\rm up})}{Nb}\left(1 + \frac{4\gamma^2 L^2 \omega_{\rm dwn}}{\alpha_{\rm dwn}}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\rm up}}{N}\right)\right).
\end{aligned}$$

We invert the double-sum, it leads to:

$$\begin{aligned}
\mathbb{E}\left[\|w_k - w_*\|^2\right] &\leq \|w_0 - w_*\|^2 - \gamma \sum_{j=1}^{k} \mathbb{E}\left[F(w_{j-1}) - F(w_*)\right] \\
&\quad - \frac{\gamma}{2L} \mathbb{E}\left[\|\nabla F(w_{k-1})\|^2\right] \\
&\quad + \frac{2}{\alpha_{\rm dwn}} \times 2\gamma^4 L^2 \omega_{\rm dwn} \left(\frac{1}{\mathbf{C}} + \frac{\omega_{\rm up}}{N}\right)\left(\frac{1}{\alpha_{\rm dwn}} + \frac{\omega_{\rm up}}{N}\right) \mathbb{E}\left[\|\nabla F(w_{-1})\|^2\right] \\
&\quad + \sum_{j=1}^{k-1} \left(\frac{2}{\alpha_{\rm dwn}} \times 2\gamma^4 L^2 \omega_{\rm dwn} \left(\frac{1}{\mathbf{C}} + \frac{\omega_{\rm up}}{N}\right)\left(\frac{1}{\alpha_{\rm dwn}} + \frac{\omega_{\rm up}}{N}\right) - \frac{\gamma}{2L}\right) \mathbb{E}\left[\|\nabla F(w_{j-1})\|^2\right] \\
&\quad + \frac{\gamma^2\sigma^2(1+\omega_{\rm up})}{Nb}\left(1 + \frac{4\gamma^2 L^2 \omega_{\rm dwn}}{\alpha_{\rm dwn}}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\rm up}}{N}\right)\right) \times k\,.
\end{aligned}$$

Now, we consider that $\frac{4\omega_{\rm dwn}\gamma^4 L^2}{\alpha_{\rm dwn}}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\rm up}}{N}\right)\left(\frac{1}{\alpha_{\rm dwn}} + \frac{\omega_{\rm up}}{N}\right) < \frac{\gamma}{2L}$, thus we have:

$$\frac{\gamma}{k}\sum_{t=1}^{k} \mathbb{E}\left[F(w_{t-1}) - F(w_*)\right] \leq \frac{\|w_0 - w_*\|^2}{k} + \frac{\gamma^2\sigma^2(1+\omega_{\rm up})}{Nb}\left(1 + \frac{4\gamma^2 L^2 \omega_{\rm dwn}}{\alpha_{\rm dwn}}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\rm up}}{N}\right)\right).$$

Finally, by Jensen, for any K in $\mathbb{N}^*$, taking γ such that:

$$\gamma L \leq \min\left\{\frac{1}{8\omega_{\rm dwn}\sqrt{\frac{1}{\mathbf{C}} + \frac{\omega_{\rm up}}{N}}}, \frac{1}{2\left(1+\frac{\omega_{\rm up}}{N}\right)}, \frac{1}{\sqrt[3]{\frac{8\omega_{\rm dwn}}{\alpha_{\rm dwn}}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\rm up}}{N}\right)\left(\frac{1}{\alpha_{\rm dwn}} + \frac{\omega_{\rm up}}{N}\right)}}\right\}$$

and with $\alpha_{\rm dwn} \leq \frac{1}{8\omega_{\rm dwn}}$, we recover Theorem 5:

$$\mathbb{E}\left[F(\bar{w}_K) - F(w_*)\right] \leq \frac{\|w_0 - w_*\|^2}{\gamma K} + \frac{\gamma\sigma^2\Phi^{\rm Rd}(\gamma)}{Nb},$$

denoting $\Phi^{\rm Rd}(\gamma) = (1+\omega_{\rm up})\left(1 + \frac{4\gamma^2 L^2 \omega_{\rm dwn}}{K}\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\rm up}}{N}\right)\right)$.

□

G Adataptation to the heterogeneous scenario

In this section, we give the complete proof of Theorems 1 and 2 in the case of heterogeneous workers.

> We choose to not merge the proofs in the homogeneous and heterogeneous cases. This is to avoid the technicalities associated with the heterogeneity and the uplink compression (that have been extensively studied in previous works [34, 18, 28, 36]) in the proof of our main results which aim at alleviating the impact of downlink compression. We thus propose two proofs that can be read almost independently in order to make proof-checking easier. We stress that the result in the homogeneous setting is not exactly a consequence of the heterogeneous case (the constants are degraded in the heterogeneous framework) but merging the proofs is ultimately possible.

Appendix G.1 first presents some lemmas from [36] required to handle the additional uplink memory. Lemma S6 (resp. Lemma S7) corresponds to Lemma S5 (resp. Lemma S7) evaluated at point $\widehat{w}_{k-1}$; and Lemma S8 corresponds to Lemma S13. Secondly, Appendix G.2 gives the demonstration of MCM. We denote $\Phi^{\text{Heterog}}(\gamma) := (1+8\omega_{\text{up}})\left(1+\frac{8\gamma L\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\right)$ and $\gamma_{\max}^{\text{Heterog}}$ such that:

$$\gamma_{\max}^{\text{Heterog}} L \leq \min\left\{\gamma_{\max}, \frac{1}{16\frac{\omega_{\text{up}}}{N}}, \frac{1}{8\sqrt{2\frac{\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\cdot\frac{\omega_{\text{up}}}{N}}}\right\}.$$

We make the following assumption on the heterogeneity.

Assumption 6 (Bounded gradient at w_*)**.** *There is a constant B in $\mathbb{R}_+$, s.t.: $\frac{1}{N}\sum_{i=0}^{N}\|\nabla F_i(w_*)\|^2 = B^2$. And we denote for all i in $[\![1,N]\!]$, $h_*^i = \nabla F_i(w_*)$.*

G.1 Control of the uplink memory

In this section we give the theorems that are required by the uplink memory.

Lemma S6 (Bounding the compressed term)**.** *The squared norm of the compressed term sent by each node to the central server can be bounded as following:*

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

Lemma S7 (Noise over local gradients)**.** *Let $k\in N^*$ and $i\in[\![1,N]\!]$. The noise in the stochastic gradients as defined in Assumptions 4 and 6 can be controlled as following:*

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

Lemma S8 (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:*

$$\begin{aligned}\mathbb{E}\left[\Xi_k\mid w_{k-1}\right] \leq{}& (1-\alpha_{\text{up}})\mathbb{E}\left[\Xi_{k-1}\mid w_{k-1}\right]\\ &+\frac{2\alpha_{\text{up}}L}{N}\mathbb{E}\left[\langle\,\nabla F(\widehat{w}_{k-1}),\,\widehat{w}_{k-1}-w_*\,\rangle\mid w_{k-1}\right]\\ &+\frac{2}{N}\frac{\sigma^2}{b}\alpha_{\text{up}}\,.\end{aligned}$$

Lemma S9 (Squared-norm of stochastic gradients)**.** *For any k in $\mathbb{N}^*$, the squared-norm of gradients can be bounded a.s.:*

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{g}_k\|^2\,\middle|\,\widehat{w}_{k-1}\right] \leq{}& \left(1+\frac{4\omega_{\text{up}}}{N}\right)L\mathbb{E}\left[\langle\,\nabla F(\widehat{w}_{k-1}),\,\widehat{w}_{k-1}-w_*\,\rangle\mid\widehat{w}_{k-1}\right]\\ &+2\omega_{\text{up}}\mathbb{E}\left[\Xi_{k-1}\mid\widehat{w}_{k-1}\right]+\frac{\sigma^2}{Nb}(1+4\omega_{\text{up}})\,,\\ \mathbb{E}\left[\|\widetilde{g}_k-\nabla F(\widehat{w}_{k-1})\|^2\,\middle|\,\widehat{w}_{k-1}\right] \leq{}& \frac{4\omega_{\text{up}}L}{N}\mathbb{E}\left[\langle\,\nabla F(\widehat{w}_{k-1}),\,\widehat{w}_{k-1}-w_*\,\rangle\mid\widehat{w}_{k-1}\right]\\ &+2\omega_{\text{up}}\mathbb{E}\left[\Xi_{k-1}\mid\widehat{w}_{k-1}\right]+\frac{\sigma^2}{Nb}(1+4\omega_{\text{up}})\,,\end{aligned}$$

Lemma S9 extends Lemma S3.

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

$$\mathbb{E}\left[\|\widetilde{g}_k\|^2 \;\middle|\; \widehat{w}_{k-1}\right] = \|\nabla F(\widehat{w}_{k-1})\|^2 + \mathbb{E}\left[\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; \widehat{w}_{k-1}\right]$$
$$\leq L\mathbb{E}\left[\langle\, \nabla F(\widehat{w}_{k-1}),\, \widehat{w}_{k-1} - w_* \,\rangle \mid \widehat{w}_{k-1}\right] + \mathbb{E}\left[\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; \widehat{w}_{k-1}\right]$$

Secondly:

$$\mathbb{E}\left[\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; \widehat{w}_{k-1}\right]$$
$$= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \left(\widehat{\Delta}_{k-1}^i + h_{k-1}^i - \nabla F_i(\widehat{w}_{k-1})\right)\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right]$$
$$= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \left(\widehat{\Delta}_{k-1}^i + h_{k-1}^i - g_k^i(\widehat{w}_{k-1}) + g_k^i(\widehat{w}_{k-1}) - \nabla F_i(\widehat{w}_{k-1})\right)\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right]$$
$$= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \widehat{\Delta}_{k-1}^i - \Delta_{k-1}^i\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right] + \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N g_k^i(\widehat{w}_{k-1}) - \nabla F_i(\widehat{w}_{k-1})\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right],$$

the inner product being null.

Next, expanding the squared norm again, and because the two sums of inner products are null as the stochastic oracle and uplink compressions are independent:

$$\mathbb{E}\left[\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; \widehat{w}_{k-1}\right] = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|\widehat{\Delta}_{k-1}^i - \Delta_{k-1}^i\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right]$$
$$+ \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i(\widehat{w}_{k-1}) - \nabla F_i(\widehat{w}_{k-1})\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right].$$

Then, for any i in $[\![1, N]\!]$ as $\mathbb{E}\left[\left\|\widehat{\Delta}_{k-1}^i - \Delta_{k-1}^i\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|\widehat{\Delta}_{k-1}^i - \Delta_{k-1}^i\right\|^2 \;\middle|\; g_k^i\right] \;\middle|\; \widehat{w}_{k-1}\right]$, and using Assumption 1 we have:

$$\mathbb{E}\left[\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; \widehat{w}_{k-1}\right] \leq \frac{\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|\Delta_{k-1}^i\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right]$$
$$+ \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k^i(\widehat{w}_{k-1}) - \nabla F_i(\widehat{w}_{k-1})\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right].$$

Furthermore with Lemma S6 and Assumption 4:

$$\mathbb{E}\left[\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; \widehat{w}_{k-1}\right] \leq \frac{2\omega_{\mathrm{up}}}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|g_k(\widehat{w}_{k-1}) - h_*^i\right\|^2 \;\middle|\; \widehat{w}_{k-1}\right]$$
$$+ 2\omega_{\mathrm{up}}\mathbb{E}\left[\Xi_{k-1} \mid \widehat{w}_{k-1}\right] + \frac{\sigma^2}{Nb}.$$

And finally with Lemma S7:

$$\mathbb{E}\left[\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\|^2 \;\middle|\; \widehat{w}_{k-1}\right] \leq \frac{4\omega_{\mathrm{up}}L}{N}\mathbb{E}\left[\langle\, \nabla F(\widehat{w}_{k-1}),\, \widehat{w}_{k-1} - w_* \,\rangle \mid \widehat{w}_{k-1}\right] + \frac{4\omega_{\mathrm{up}}\sigma^2}{Nb}$$
$$+ 2\omega_{\mathrm{up}}\mathbb{E}\left[\Xi_{k-1} \mid \widehat{w}_{k-1}\right] + \frac{\sigma^2}{Nb},$$

from which we derive the two inequalities of the lemma.

□

G.2 Proofs for MCM

In this section, we provide the demonstration of Theorems 1 and 2 in the convex and strongly-convex cases with heterogeneous workers.

G.2.1 Control of the Variance of the local model for MCM

In this section, the aim is to control the variance of the local model for MCM but in the setting of heterogeneous worker, as done previously in Theorem S8.

Theorem S14. *Consider the MCM update as in eq. (2). Under Assumptions 1, 2 and 4, if* $\gamma \leq 1/(8\omega_{\rm dwn}L)$ *and* $\alpha_{\rm dwn} \leq 1/(8\omega_{\rm dwn})$*, then for all* k *in* $\mathbb{N}$*:*

$$\begin{aligned}\mathbb{E}\left[\Upsilon_k \mid w_{k-1}\right] \leq \left(1 - \frac{\alpha_{\rm dwn}}{2}\right)\Upsilon_{k-1} \\ + 2\gamma^2 L\left(\frac{1}{\alpha_{\rm dwn}} + \frac{4\omega_{\rm up}}{N}\right)\mathbb{E}\left[\langle\, \nabla F(\widehat{w}_{k-1}),\, \widehat{w}_{k-1} - w_* \,\rangle \mid \widehat{w}_{k-1}\right] \\ + 4\gamma^2\omega_{\rm up}\mathbb{E}\left[\Xi_{k-1} \mid \widehat{w}_{k-1}\right] + \frac{2\gamma^2\sigma^2(1+4\omega_{\rm up})}{Nb}\,.\end{aligned}$$

Proof. Let k in $\mathbb{N}$, we recall that by definition:

$$\begin{cases} \Omega_k = w_k - H_{k-1} \\ \widehat{\Omega}_k = \mathcal{C}_{\rm dwn}(\Omega_k) \\ \widehat{w}_k = \widehat{\Omega}_k + H_{k-1}\,. \end{cases}$$

We start the proof by performing a bias-variance decomposition, and exactly like in the proof of Theorem S8, we obtain:

$$\|\Omega_k\|^2 = \text{Bias}^2 + 2\gamma^2\text{Var}_{12} + 2\gamma^2\text{Var}_{12} + 2\alpha_{\rm dwn}^2\text{Var}_2$$

We first have:

$$\text{Var}_{11} = \mathbb{E}\left[\left\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\right\|^2 \,\middle|\, w_{k-1}\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|\widetilde{g}_k - \nabla F(\widehat{w}_{k-1})\right\|^2 \,\middle|\, \widehat{w}_{k-1}\right] \,\middle|\, w_{k-1}\right]\,,$$

so, we can use Lemma S9:

$$\text{Var}_{11} = \frac{4\omega_{\rm up}L}{N}\mathbb{E}\left[\langle\, \nabla F(\widehat{w}_{k-1}),\, \widehat{w}_{k-1} - w_* \,\rangle \mid \widehat{w}_{k-1}\right] + 2\omega_{\rm up}\mathbb{E}\left[\Xi_{k-1} \mid \widehat{w}_{k-1}\right] + \frac{\sigma^2}{Nb}(1+4\omega_{\rm up})\,.$$

The other terms are exactly as before in Theorem S8:

$$\begin{cases} \text{Var}_{12} \leq L^2\omega_{\rm dwn}\Upsilon_{k-1} \\ \text{Var}_2 \leq \omega_{\rm dwn}\Upsilon_{k-1} \\ \text{Bias}^2 \leq (1-\alpha_{\rm dwn})\Upsilon_{k-1} + \gamma^2 L(1+\frac{1}{\alpha_{\rm dwn}})\mathbb{E}\left[\langle\, \nabla F(\widehat{w}_{k-1}),\, \widehat{w}_{k-1} - w_* \,\rangle \mid \widehat{w}_{k-1}\right]\,. \end{cases}$$

At the end:

$$\begin{aligned}\mathbb{E}\left[\Upsilon_k \mid w_{k-1}\right] \leq (1-\alpha_{\rm dwn})\Upsilon_{k-1} + \gamma^2 L(1+\frac{1}{\alpha_{\rm dwn}})\mathbb{E}\left[\langle\, \nabla F(\widehat{w}_{k-1}),\, \widehat{w}_{k-1} - w_* \,\rangle \mid \widehat{w}_{k-1}\right] \\ + \frac{8\omega_{\rm up}\gamma^2 L}{N}\mathbb{E}\left[\langle\, \nabla F(\widehat{w}_{k-1}),\, \widehat{w}_{k-1} - w_* \,\rangle \mid \widehat{w}_{k-1}\right] \\ + 4\gamma^2\omega_{\rm up}\mathbb{E}\left[\Xi_{k-1} \mid \widehat{w}_{k-1}\right] + \frac{2\gamma^2\sigma^2}{Nb}(1+4\omega_{\rm up}) \\ + 2\gamma^2 L^2\omega_{\rm dwn}\Upsilon_{k-1} + 2\alpha_{\rm dwn}^2\omega_{\rm dwn}\Upsilon_{k-1}\,,\end{aligned}$$

which is equivalent to:

$$\mathbb{E}\left[\Upsilon_k \mid w_{k-1}\right] \leq \left(1-\alpha_{\mathrm{dwn}}+2\gamma^2L^2\omega_{\mathrm{dwn}}+2\alpha_{\mathrm{dwn}}^2\omega_{\mathrm{dwn}}\right)\left\|w_{k-1}-\Xi_{k-2}\right\|^2 + \gamma^2L\left(1+\frac{1}{\alpha_{\mathrm{dwn}}}+\frac{8\omega_{\mathrm{up}}}{N}\right)\mathbb{E}\left[\left\langle \nabla F(\widehat{w}_{k-1}),\ \widehat{w}_{k-1}-w_*\right\rangle \mid \widehat{w}_{k-1}\right] + 4\gamma^2\omega_{\mathrm{up}}\mathbb{E}\left[\Xi_{k-1}\mid \widehat{w}_{k-1}\right]+\frac{2\gamma^2\sigma^2(1+4\omega_{\mathrm{up}})}{Nb}\,.$$

Next, we require as in Theorem S8:

$$\begin{cases} 2\alpha_{\mathrm{dwn}}^2\omega_{\mathrm{dwn}} \leq \frac{1}{4}\alpha_{\mathrm{dwn}} \iff \alpha_{\mathrm{dwn}} \leq \frac{1}{8\omega_{\mathrm{dwn}}}\,, \\ 2\gamma^2L^2\omega_{\mathrm{dwn}} \leq \frac{1}{4}\alpha_{\mathrm{dwn}} = \frac{1}{32\omega_{\mathrm{dwn}}}\,, \text{ by taking } \alpha_{\mathrm{dwn}} = \frac{1}{8\omega_{\mathrm{dwn}}} \iff \gamma \leq \frac{1}{8\omega_{\mathrm{dwn}}L}\,, \\ 1+\frac{1}{\alpha_{\mathrm{dwn}}} \leq \frac{2}{\alpha_{\mathrm{dwn}}} \text{ which is not restrictive if } \omega_{\mathrm{dwn}} \geq 1, \end{cases}$$

and it leads to the final result taking unconditional expectation. □

G.2.2 Convex case

Theorem S15 (Convergence of MCM in the heterogeneous and convex case)**.** *Under Assumptions 1 to 4 with $\mu = 0$ (convex case), for learning rates $\alpha_{\mathrm{dwn}} \leq \frac{1}{8\omega_{\mathrm{dwn}}}$ and $\alpha_{\mathrm{up}}(1+\omega_{\mathrm{up}}) \leq 1$, taking a step size s.t. $\gamma \leq \gamma_{\max}^{\mathrm{Heterog}}$, for any k in $\mathbb{N}$, defining:*

$$V_k := \mathbb{E}\left[\left\|w_k - w_*\right\|^2\right] + \gamma^2C_1\mathbb{E}\left[\Xi_k\right] + \gamma LC_2\mathbb{E}\left[\Upsilon_k\right]\,,$$

with $C_1 = 2\omega_{\mathrm{up}}(1+8\gamma L\omega_{\mathrm{dwn}}/\alpha_{\mathrm{dwn}})/\alpha_{\mathrm{up}}$, $C_2 = 4\gamma L\omega_{\mathrm{dwn}}/\alpha_{\mathrm{dwn}}$, we have:

$$V_k \leq V_{k-1} - \gamma\mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] + \frac{\gamma^2\sigma^2\Phi^{\mathrm{Heterog}}(\gamma)}{Nb}\,.$$

Proof. We denote for k in $\mathbb{N}^*$ $\widetilde{g}_k = \frac{1}{N}\sum_{i=1}^N \widehat{\Delta}_{k-1}^i + h_{k-1}^i$ with $\Delta_{k-1}^i = g_k^i(\widehat{w}_{k-1}) - h_{k-1}^i$, and $\Xi_k = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|h_k^i - \nabla F_i(w_*)\right\|^2 \,\middle|\, \widehat{w}_{k-1}\right]$.

Let k in $\mathbb{N}^*$, by definition:

$$\left\|w_k - w_*\right\|^2 \leq \left\|w_{k-1} - w_*\right\|^2 - 2\gamma\left\langle \widetilde{g}_k,\ w_{k-1} - w_*\right\rangle + \gamma^2\left\|\widetilde{g}_k\right\|^2\,.$$

Next, we expend the inner product as following:

$$\left\|w_k - w_*\right\|^2 \leq \left\|w_{k-1} - w_*\right\|^2 - 2\gamma\left\langle \widetilde{g}_k,\ \widehat{w}_{k-1} - w_*\right\rangle - 2\gamma\left\langle \widetilde{g}_k,\ w_{k-1} - \widehat{w}_{k-1}\right\rangle + \gamma^2\left\|\widetilde{g}_k\right\|^2\,.$$

Taking expectation conditionally to w_{k-1}, and using $\mathbb{E}\left[\widetilde{g}_k \mid w_{k-1}\right] = \mathbb{E}\left[\mathbb{E}\left[\widetilde{g}_k \mid \widehat{w}_{k-1}\right] \mid w_{k-1}\right] = \mathbb{E}\left[\nabla F(\widehat{w}_{k-1}) \mid w_{k-1}\right]$, we obtain:

$$\mathbb{E}\left[\left\|w_k - w_*\right\|^2 \,\middle|\, w_{k-1}\right] \leq \left\|w_{k-1} - w_*\right\|^2 - \mathbb{E}\left[2\gamma\left\langle \nabla F(\widehat{w}_{k-1}),\ \widehat{w}_{k-1} - w_*\right\rangle \mid w_{k-1}\right] - 2\gamma\mathbb{E}\left[\left\langle \nabla F(\widehat{w}_{k-1}),\ w_{k-1} - \widehat{w}_{k-1}\right\rangle \mid w_{k-1}\right] + \gamma^2\mathbb{E}\left[\left\|\widetilde{g}_k\right\|^2 \,\middle|\, w_{k-1}\right]\,.$$

Then invoking Lemma S3 to upper bound the squared norm of the stochastic gradients, and noticing that $\mathbb{E}\left[\left\langle \nabla F(w_{k-1}),\ \widehat{w}_{k-1} - w_{k-1}\right\rangle \mid w_{k-1}\right] = 0$ leads to:

$$\begin{aligned}\mathbb{E}\left[\left\|w_k - w_*\right\|^2 \,\middle|\, w_{k-1}\right] \leq{}& \left\|w_{k-1} - w_*\right\|^2 - 2\gamma\mathbb{E}\left[\left\langle \nabla F(\widehat{w}_{k-1}),\ \widehat{w}_{k-1} - w_*\right\rangle \mid w_{k-1}\right] \\ &- 2\gamma\mathbb{E}\left[\left\langle \nabla F(\widehat{w}_{k-1}) - \nabla F(w_{k-1}),\ w_{k-1} - \widehat{w}_{k-1}\right\rangle \mid w_{k-1}\right] \\ &+ \gamma^2\left(\left(1+\frac{4\omega_{\mathrm{up}}}{N}\right)L\mathbb{E}\left[\left\langle \nabla F(\widehat{w}_{k-1}),\ \widehat{w}_{k-1} - w_*\right\rangle \mid \widehat{w}_{k-1}\right]\right. \\ &\left. +2\omega_{\mathrm{up}}\mathbb{E}\left[\Xi_{k-1}\mid \widehat{w}_{k-1}\right] + \frac{\sigma^2}{Nb}(1+4\omega_{\mathrm{up}})\right)\,.\end{aligned} \tag{S25}$$

Now using Cauchy-Schwarz inequality (eq. (S5)) and smoothness:

$$\begin{aligned}
&- \mathbb{E}\left[2\gamma \left\langle \nabla F(\widehat{w}_{k-1}) - \nabla F(w_{k-1}),\, w_{k-1} - \widehat{w}_{k-1} \right\rangle \mid w_{k-1}\right] \\
&\qquad = 2\gamma \mathbb{E}\left[\left\langle \nabla F(\widehat{w}_{k-1}) - \nabla F(w_{k-1}),\, \widehat{w}_{k-1} - w_{k-1} \right\rangle \mid w_{k-1}\right] \\
&\qquad \leq 2\gamma L \mathbb{E}\left[\left\|\widehat{w}_{k-1} - w_{k-1}\right\|^2 \;\middle|\; w_{k-1}\right],
\end{aligned}$$

and thus:

$$\begin{aligned}
\mathbb{E}\left[\left\|w_k - w_*\right\|^2 \;\middle|\; w_{k-1}\right] \leq \left\|w_{k-1} - w_*\right\|^2 &- 2\gamma \mathbb{E}\left[\left\langle \nabla F(\widehat{w}_{k-1}),\, \widehat{w}_{k-1} - w_* \right\rangle \mid w_{k-1}\right] \\
&+ 2\gamma L \mathbb{E}\left[\left\|\widehat{w}_{k-1} - w_{k-1}\right\|^2 \;\middle|\; w_{k-1}\right] \\
&+ \left(1 + \frac{4\omega_{\mathrm{up}}}{N}\right) \gamma^2 L \mathbb{E}\left[\left\langle \nabla F(\widehat{w}_{k-1}),\, \widehat{w}_{k-1} - w_* \right\rangle \mid \widehat{w}_{k-1}\right] \\
&+ 2\omega_{\mathrm{up}} \gamma^2 \mathbb{E}\left[\Xi_{k-1} \mid \widehat{w}_{k-1}\right] + \frac{\sigma^2 \gamma^2}{Nb}(1 + 4\omega_{\mathrm{up}}).
\end{aligned}$$

As $\gamma \leq \dfrac{1}{2L\left(1 + \frac{\omega_{\mathrm{up}}}{N}\right)}$, and thus $\left(1 - \frac{\gamma L}{2}\left(1 + \frac{4\omega_{\mathrm{up}}}{N}\right)\right) \geq 1/2$; this allows to simplify the coefficient of the scalar product:

$$\begin{aligned}
\mathbb{E}\left[\left\|w_k - w_*\right\|^2 \;\middle|\; w_{k-1}\right] \leq \left\|w_{k-1} - w_*\right\|^2 &- \gamma \mathbb{E}\left[\left\langle \nabla F(\widehat{w}_{k-1}),\, \widehat{w}_{k-1} - w_* \right\rangle \mid w_{k-1}\right] \\
&+ 2\gamma L \mathbb{E}\left[\left\|\widehat{w}_{k-1} - w_{k-1}\right\|^2 \;\middle|\; w_{k-1}\right] \\
&+ 2\omega_{\mathrm{up}} \gamma^2 \mathbb{E}\left[\Xi_{k-1} \mid \widehat{w}_{k-1}\right] + \frac{\sigma^2 \gamma^2}{Nb}(1 + 4\omega_{\mathrm{up}}).
\end{aligned} \tag{S26}$$

With Lemma S8, we have :

$$\begin{aligned}
\mathbb{E}\left[\Xi_k \mid w_{k-1}\right] \leq (1 - \alpha_{\mathrm{up}}) &\mathbb{E}\left[\Xi_{k-1} \mid w_{k-1}\right] \\
&+ \frac{2\alpha_{\mathrm{up}} L}{N} \mathbb{E}\left[\left\langle \nabla F(\widehat{w}_{k-1}),\, \widehat{w}_{k-1} - w_* \right\rangle \mid w_{k-1}\right] \\
&+ \frac{2}{N}\frac{\sigma^2}{b}\alpha_{\mathrm{up}}.
\end{aligned} \tag{S27}$$

and Theorem S14 gives:

$$\begin{aligned}
\mathbb{E}\left[\Upsilon_k \mid w_{k-1}\right] \leq \left(1 - \frac{\alpha_{\mathrm{dwn}}}{2}\right) &\Upsilon_{k-1} \\
&+ 2\gamma^2 L \left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{4\omega_{\mathrm{up}}}{N}\right) \mathbb{E}\left[\left\langle \nabla F(\widehat{w}_{k-1}),\, \widehat{w}_{k-1} - w_* \right\rangle \mid \widehat{w}_{k-1}\right] \\
&+ 4\gamma^2 \omega_{\mathrm{up}} \mathbb{E}\left[\Xi_{k-1} \mid \widehat{w}_{k-1}\right] + \frac{2\gamma^2 \sigma^2 (1 + 4\omega_{\mathrm{up}})}{Nb}.
\end{aligned} \tag{S28}$$

We take the full expectation (without conditioning) and we set:

$$V_k := \mathbb{E}\left[\left\|w_k - w_*\right\|^2\right] + \gamma^2 C_1 \mathbb{E}\left[\Xi_k\right] + \gamma L C_2 \mathbb{E}\left[\Upsilon_k\right],$$

with $C_1 = 2\omega_{\mathrm{up}}(1 + 8\gamma L \omega_{\mathrm{dwn}}/\alpha_{\mathrm{dwn}})/\alpha_{\mathrm{up}}$ and $C_2 = 4\omega_{\mathrm{dwn}}/\alpha_{\mathrm{dwn}}$.

We combine previous equations as follows (S26) $+ \gamma^2 C_1$(S27) $+ C_2$(S28):

$$\begin{aligned}
\mathbb{E}\left[\|w_k - w_*\|^2\right] + \gamma^2 C_1 \mathbb{E}\left[\Xi_k\right] + \gamma L C_2 \mathbb{E}\left[\Upsilon_k\right] &\leq \|w_{k-1} - w_*\|^2 \\
&- \gamma\left(1 - \gamma L\left(\left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{4\omega_{\mathrm{up}}}{N}\right)\gamma L C_2 + \frac{\alpha_{\mathrm{up}} C_1}{N}\right)\right)\mathbb{E}\left[\langle\, \nabla F(\widehat{w}_{k-1}),\ \widehat{w}_{k-1} - w_*\,\rangle\right] \\
&+ \left(2\omega_{\mathrm{up}}(1 + 2\gamma L C_2) + (1 - \alpha_{\mathrm{up}})C_1\right)\gamma^2 \mathbb{E}\left[\Xi_{k-1}\right] \\
&+ \left(2\gamma L \omega_{\mathrm{dwn}} + \left(1 - \frac{\alpha_{\mathrm{dwn}}}{2}\right)\gamma L C_2\right)\mathbb{E}\left[\Upsilon_{k-1}\right] \\
&+ \frac{\gamma^2 \sigma^2}{Nb}\left((1 + 4\omega_{\mathrm{up}})(1 + 2\gamma L C_2) + 2\alpha_{\mathrm{up}} C_1\right)\,,
\end{aligned} \tag{S29}$$

We first observe that:

$$2\gamma L \omega_{\mathrm{dwn}} + \left(1 - \frac{\alpha_{\mathrm{dwn}}}{2}\right)\gamma L C_2 \leq \gamma L C_2 \iff C_2 \geq \frac{4\omega_{\mathrm{dwn}}}{\alpha_{\mathrm{dwn}}}\,, \quad \text{which is true by definition of } C_2.$$

Secondly, ensuring that the factor multiplying $\mathbb{E}\left[\Xi_{k-1}\right]$ on the right hand side is smaller than $\gamma^2 C_1$ requires:

$$\begin{aligned}
& 2\omega_{\mathrm{up}}(1 + 2\gamma L C_2) + (1 - \alpha_{\mathrm{up}})C_1 \leq C_1 \\
\implies \quad & C_1 \geq \frac{2\omega_{\mathrm{up}}(1 + 8\gamma L \omega_{\mathrm{dwn}}/\alpha_{\mathrm{dwn}})}{\alpha_{\mathrm{up}}} \quad \text{because } C_2 = 4\omega_{\mathrm{dwn}}/\alpha_{\mathrm{dwn}}.
\end{aligned}$$

Finally, we have that $1 - \gamma L\left(\left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{4\omega_{\mathrm{up}}}{N}\right)\gamma L C_2 + \frac{\alpha_{\mathrm{up}} C_1}{N}\right) \geq \frac{1}{2}$, if we take γ such that:

$$\begin{cases}
\left(\frac{1}{\alpha_{\mathrm{dwn}}} + \frac{4\omega_{\mathrm{up}}}{N}\right)(\gamma L)^2 C_2 \leq 1/4 \implies \gamma \leq \dfrac{1}{4L\sqrt{\dfrac{\omega_{\mathrm{dwn}}}{\alpha_{\mathrm{dwn}}}\left(\dfrac{1}{\alpha_{\mathrm{dwn}}} + \dfrac{4\omega_{\mathrm{up}}}{N}\right)}} \\
\dfrac{\gamma L \alpha_{\mathrm{up}} C_1}{N} \leq 1/4 \iff \dfrac{2\gamma L \omega_{\mathrm{up}}}{N}\left(1 + 8\gamma L \omega_{\mathrm{dwn}}/\alpha_{\mathrm{dwn}}\right) \leq 1/4
\end{cases}$$

We rewrite the second condition as follows:

$$\begin{cases}
16(\gamma L)^2 \dfrac{\omega_{\mathrm{up}}\omega_{\mathrm{dwn}}}{\alpha_{\mathrm{dwn}} N} \leq 1/8 \iff \gamma \leq \dfrac{1}{8L\sqrt{2\dfrac{\omega_{\mathrm{dwn}}}{\alpha_{\mathrm{dwn}}} \cdot \dfrac{\omega_{\mathrm{up}}}{N}}} \\
\dfrac{2\gamma L \omega_{\mathrm{up}}}{N \alpha_{\mathrm{up}}} \leq 1/8 \iff \gamma \leq \dfrac{N}{16 L \omega_{\mathrm{up}}}\,.
\end{cases}$$

Applying convexity, we derive:

$$V_k \leq V_{k-1} - \gamma \mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] + \frac{\gamma^2 \sigma^2 \Phi^{\mathrm{Heterog}}(\gamma)}{Nb}\,,$$

with $\Phi^{\mathrm{Heterog}}(\gamma) = (1 + 8\omega_{\mathrm{up}})\left(1 + \frac{8\gamma L \omega_{\mathrm{dwn}}}{\alpha_{\mathrm{dwn}}}\right)$. Invoking Jensen inequality (S7) leads to $\mathbb{E}\left[F(\widehat{w}_{k-1})\right] \geq \mathbb{E}\left[F(w_{k-1})\right]$, and we finally obtain:

$$V_k \leq V_{k-1} - \gamma \mathbb{E}\left[F(w_{k-1}) - F(w_*)\right] + \frac{\gamma^2 \sigma^2 \Phi^{\mathrm{Heterog}}(\gamma)}{Nb}\,.$$

□

G.2.3 Strongly-convex case

Theorem S16 (Convergence of MCM in the heterogeneous and strongly-convex case)**.** *Under Assumptions 1 to 4 with $\mu = 0$ (convex case), for learning rates $\alpha_{\rm dwn} \leq \frac{1}{8\omega_{\rm dwn}}$ and $\alpha_{\rm up}(1+\omega_{\rm up}) \leq 1$, for any sequence $(\gamma_k)_{k\in\mathbb{N}} \leq \gamma_{\max}^{\rm Heterog}$, for any k in $\mathbb{N}$, defining:*

$$V_k := \mathbb{E}\left[\|w_k - w_*\|^2\right] + \gamma_k^2 C_1 \mathbb{E}\left[\Xi_k\right] + \gamma_k L C_2 \mathbb{E}\left[\Upsilon_k\right] ,$$

with $C_1 = 2\omega_{\rm up}(1 + 8\gamma L \omega_{\rm dwn}/\alpha_{\rm dwn})/\alpha_{\rm up}$, $C_2 = 4\gamma L \omega_{\rm dwn}/\alpha_{\rm dwn}$, we have:

$$V_k \leq (1-\gamma_k\mu)V_{k-1} - \gamma \mathbb{E}\left[F(\widehat{w}_{k-1}) - F(w_*)\right] + \frac{\gamma^2\sigma^2\Phi^{\rm Heterog}(\gamma)}{Nb} .$$

Proof. Let k in $\mathbb{N}^*$, the proof starts like the one for MCM in the convex case with heterogeneous worker, and we start from eq. (S26) but we consider a variable step size $\gamma_k = 2/(\mu(k+1)+\widetilde{L})$ that depends of the iteration k in $\mathbb{N}$.

We consider this following Lyapunov function:

$$V_k = \mathbb{E}\left[\|w_k - w_*\|^2\right] + \gamma_k^2 C_1 \mathbb{E}\left[\Xi_k\right] + \gamma_k L C_2 \mathbb{E}\left[\Upsilon_k\right] ,$$

with $C_1 = 2\omega_{\rm up}(1 + 8\gamma_k L \omega_{\rm dwn}/\alpha_{\rm dwn})/\alpha_{\rm up}$ and $C_2 = 4\omega_{\rm dwn}/\alpha_{\rm dwn}$.

$$\begin{aligned}
\mathbb{E}\left[\|w_k - w_*\|^2\right] + \gamma_k^2 C_1 \mathbb{E}\left[\Xi_k\right] + \gamma_k L C_2 \mathbb{E}\left[\Upsilon_k\right] &\leq \|w_{k-1} - w_*\|^2 \\
&- \gamma_k \left(1 - \gamma_k L \left(\left(\frac{1}{\alpha_{\rm dwn}} + \frac{4\omega_{\rm up}}{N}\right)\gamma_k L C_2 + \frac{\alpha_{\rm up} C_1}{N}\right)\right) \mathbb{E}\left[\langle\, \nabla F(\widehat{w}_{k-1}),\, \widehat{w}_{k-1} - w_* \,\rangle\right] \\
&+ \left(2\omega_{\rm up}(1 + 2\gamma_k L C_2) + (1-\alpha_{\rm up})C_1\right)\gamma_k^2 \mathbb{E}\left[\Xi_{k-1}\right] \\
&+ \left(2\gamma_k L \omega_{\rm dwn} + \left(1 - \frac{\alpha_{\rm dwn}}{2}\right)\gamma_k L C_2\right)\mathbb{E}\left[\Upsilon_{k-1}\right] \\
&+ \frac{\gamma_k^2\sigma^2}{Nb}\left((1+4\omega_{\rm up})(1+2\gamma_k L C_2) + 2\alpha_{\rm up} C_1\right) ,
\end{aligned}$$

To ensure a $(1-\gamma\mu)$-convergence we first choose $\left(1 - \frac{\alpha_{\rm dwn}}{2} + \frac{2\omega_{\rm dwn}}{C_2}\right)\gamma_k L C_2 \leq (1 - \gamma_k\mu)\gamma_{k-1} L C_2$ i.e $1 - \frac{\alpha_{\rm dwn}}{2} + \frac{2\omega_{\rm dwn}}{C_2} \leq \frac{(1-\gamma_k\mu)\gamma_{k-1}}{\gamma_k}$.

We need that for all $k \in \mathbb{N}$, $\frac{1-\gamma_k\mu}{\gamma_k} \leq \frac{1}{\gamma_{k-1}}$ i.e., $1 - \gamma_k\mu \leq \frac{\gamma_k}{\gamma_{k-1}}$, but:

$$\frac{\gamma_k}{\gamma_{k-1}} = \frac{\mu k - \mu + \widetilde{L}}{\mu k + \widetilde{L}} = 1 - \frac{\mu}{\mu k + \widetilde{L}} \quad \text{and} \quad 1 - \gamma_k\mu = 1 - \frac{2\mu}{\mu k + \widetilde{L}} ,$$

and so, the inequality is always true.

Thus we must have $2\omega_{\rm dwn}/C_2 \leq \alpha_{\rm dwn}/2$ which is true by definition of C_2.

Secondly, we need:

$$\begin{aligned}
&\left(2\omega_{\rm up}(1 + 2\gamma_k L C_2) + (1-\alpha_{\rm up})C_1\right)\gamma_k^2 \leq (1-\gamma_k\mu)\gamma_{k-1}^2 C_1 \\
\iff\quad &2\omega_{\rm up}(1 + 2\gamma_k L C_2) + (1-\alpha_{\rm up})C_1 \leq \frac{\gamma_{k-1}}{\gamma_k} C_1 \quad \text{because } \frac{1-\gamma_k\mu}{\gamma_k} \leq \frac{1}{\gamma_{k-1}},
\end{aligned}$$

because $\gamma_k/\gamma_k \leq \gamma_{k-1}/\gamma_k$, it is true if we verify the following stronger condition:

$$\begin{aligned}
&2\omega_{\rm up}(1 + 2\gamma_k L C_2) + (1-\alpha_{\rm up})C_1 \leq \frac{\gamma_k}{\gamma_k} C_1 \\
&C_1 \geq \frac{2\omega_{\rm up}\left(1 + 8\gamma_k L \omega_{\rm dwn})/\alpha_{\rm dwn}\right)}{\alpha_{\rm up}} \quad \text{because } C_2 = 4\omega_{\rm dwn}/\alpha_{\rm dwn} .
\end{aligned}$$

Finally, in order to apply convexity we must verify: $1 - \gamma L \left(\left(\frac{1}{\alpha_{\rm dwn}} + \frac{4\omega_{\rm up}}{N} \right) C_2 + \frac{\alpha_{\rm up} C_1}{N} \right) \geq \frac{1}{2}$.

We take γ_k such that:

$$\begin{cases} \left(\frac{1}{\alpha_{\rm dwn}} + \frac{4\omega_{\rm up}}{N} \right) \gamma_k L C_2 \leq 1/4 \Longrightarrow \gamma_k \leq \dfrac{1}{4L\sqrt{\dfrac{\omega_{\rm dwn}}{\alpha_{\rm dwn}} \left(\dfrac{1}{\alpha_{\rm dwn}} + \dfrac{4\omega_{\rm up}}{N} \right)}} \\ \dfrac{\gamma L \alpha_{\rm up} C_1}{N} \leq 1/4 \Longleftrightarrow \dfrac{2\gamma_k L \omega_{\rm up}}{N} \left(1 + 8\gamma_k L \omega_{\rm dwn} / \alpha_{\rm dwn} \right) \leq 1/4 \end{cases}$$

We rewrite the second condition as following:

$$\begin{cases} 16(\gamma_k L)^2 \dfrac{\omega_{\rm dwn}}{\alpha_{\rm dwn} N} \leq 1/8 \Longleftrightarrow \gamma_k \leq \dfrac{1}{8L\sqrt{2\dfrac{\omega_{\rm dwn}}{\alpha_{\rm dwn}} \cdot \dfrac{\omega_{\rm up}}{N}}} \\ \dfrac{2\gamma_k L \omega_{\rm up}}{N} \leq 1/8 \Longleftrightarrow \gamma_k \leq \dfrac{1}{16L\dfrac{\omega_{\rm up}}{N}} \, . \end{cases}$$

Now, we can apply strong-convexity:

$$V_k \leq (1 - \gamma_k \mu) V_{k-1} - \gamma_k \mathbb{E} \left[F(\widehat{w}_{k-1}) - F(w_*) \right] + \frac{\gamma_k^2 \sigma^2 \Phi^{\rm Heterog}(\gamma_k)}{Nb} \, ,$$

with $\Phi^{\rm Heterog}(\gamma) = (1 + 8\omega_{\rm up}) \left(1 + \frac{8\gamma_k L \omega_{\rm dwn}}{\alpha_{\rm dwn}} \right)$.

Invoking Jensen inequality (S7) leads to $\mathbb{E}\left[F(\widehat{w}_{k-1})\right] \geq \mathbb{E}\left[F(w_{k-1})\right]$, we finally obtain:

$$V_k \leq (1 - \gamma_k \mu) V_{k-1} - \gamma_k \mathbb{E} \left[F(w_{k-1}) - F(w_*) \right] + \frac{\gamma_k^2 \sigma^2 \Phi^{\rm Heterog}(\gamma_k)}{Nb} \, .$$

□

H Neurips Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Sections 3 and 4.
   (b) Did you describe the limitations of your work? [Yes] For `Rand-MCM`, see Subsection 4.1.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 1.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See all assumptions in Section 3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See all demonstrations in Appendices D to G
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] All the code is provided on our github repository
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix B.4

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We used four dataset : cifar10, mnist, quantum and superconduct.
   (b) Did you mention the license of the assets? [No] The dataset are under the MIT licence which is a short and simple permissive license with conditions only requiring preservation of copyright and license notices. As our work is under the same licence, there is no need to remind the licence of the four used dataset.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]