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]:

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

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].

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 held 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 $w_\ast$, and denote $F_\ast = F(w_\ast)$. 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.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Compr.</th>
<th>$e_k^i$</th>
<th>$h_k^i$</th>
<th>$E_k$</th>
<th>$H_k$</th>
<th>Rand.</th>
<th>update point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qsgd [3]</td>
<td>one-way</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECQ-sgd [49]</td>
<td>one-way</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diana [34]</td>
<td>one-way</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dore [29]</td>
<td>two-way</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>degraded</td>
</tr>
<tr>
<td>Double-Squeeze [45], Dist-EF-SGD [52]</td>
<td>two-way</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>degraded</td>
</tr>
<tr>
<td>Artemis [36]</td>
<td>two-way</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>degraded</td>
</tr>
<tr>
<td>MCM</td>
<td>two-way</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>non-degraded</td>
</tr>
<tr>
<td>Rand-MCM</td>
<td>two-way</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>non-degraded</td>
</tr>
</tbody>
</table>

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 $\gamma$, SGD corresponds to:

$$w_{k+1} = w_k - \gamma \frac{1}{N} \sum_{i=1}^{N} g_{k+1}^i(w_k).$$  \hspace{1cm} (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 $C_{\text{up}}$ and $C_{\text{dwn}}$ to compress the message in each direction. Roughly speaking, the update in eq. (1) becomes:

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

However, this approach has a major drawback. The central server receives and aggregates information $\frac{1}{N} \sum_{i=1}^{N} C_{\text{up}}(g_{k+1}^i(w_k))$. 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:

$$\begin{align*}
&\left\{ \begin{array}{l}
\hat{w}_{k+1} = w_k - \gamma \frac{1}{N} \sum_{i=1}^{N} C_{\text{up}}(g_{k+1}^i(w_k)) \\
\hat{w}_{k+1} = C_{\text{dwn}}(w_{k+1})
\end{array} \right. \hspace{1cm} \text{(2)}
\end{align*}$$

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**

$$\begin{align*}
\Omega_{k+1} &= w_{k+1} - H_k, \\
\hat{w}_{k+1} &= H_k + C_{\text{dwn}}(\Omega_{k+1}) \\
H_{k+1} &= H_k + \alpha_{\text{dwn}} C_{\text{dwn}}(\Omega_{k+1}).
\end{align*}$$

**Uplink**

$$\begin{align*}
\forall i \in \{1, N\}, & \Delta_k^i = g_{k+1}^i(\hat{w}_k) - h_k^i \\
w_{k+1} &= w_k - \gamma \sum_{i=1}^{N} C_{\text{up}}(\Delta_k^i) + h_k^i \\
h_{k+1} &= h_k^i + \alpha_{\text{up}} C_{\text{up}}(\Delta_k^i).
\end{align*}$$  \hspace{1cm} (3)

**Downlink Communication.** We introduce a **downlink memory term** $(H_k)_k$, which is available on both workers and central server. The difference $\Omega_{k+1}$ 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 $\alpha_{\text{dwn}}$.

Introducing this memory mechanism is crucial to control the variance of the local model $\hat{w}_{k+1}$. 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 $\Delta_k$ between the stochastic gradient $g^k_{i+1}$ at the local model $\hat{w}_k$ (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 $\alpha_{\text{dwn}}$.

**Remark 1** (Rate $\alpha_{\text{dwn}}$). It is necessary to use $\alpha_{\text{dwn}} < 1$. 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 $\hat{w}_k - \hat{w}_{k-1}$, (i.e., $\alpha_{\text{dwn}} = 1$), 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 $(1 + \omega_{\text{dwn}})^{-1}$) 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 $C_{\text{dwn}}$, $C_{\text{up}}$ and allow the corresponding compressions levels to be distinct: those quantities are defined in Assumption 1.

**Assumption 1.** There exists constants $\omega_{\text{up}}, \omega_{\text{dwn}} \in \mathbb{R}_+^*$, such that the compression operators $C_{\text{up}}$ and $C_{\text{dwn}}$ satisfy the two following properties for all $w$ in $\mathbb{R}^d$: $E[C_{\text{up/dwn}}(w)] = w$, and $E[|C_{\text{up/dwn}}(w) - w|^2] \leq \omega_{\text{up/dwn}}|w|^2$. The higher is $\omega$, 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 $C_{\text{dwn},i}$ instead of a single one $C_{\text{dwn}}$, 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 $i$: $C_{\text{dwn},i}(w_{k+1} - H^k_i)$. 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 updating 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 $F : \mathbb{R}^d \rightarrow \mathbb{R}$. 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 $\mu$-strongly convex (or convex if $\mu = 0$), that is for all vectors $w_1, w_2$ in $\mathbb{R}^d$: $F(w_2) \geq F(w_1) + (w_2 - w_1)^T \nabla F(w_1) + \frac{\mu}{2} \| w_2 - w_1 \|^2$.
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 $\alpha_{up} = 0$. 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 $T_k = \|w_k - H_{k-1}\|^2$, and define $V_k = E[\|w_k - w_*\|^2] + 32\gamma L_{\text{up}}^2 E[T_k]$, 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 $\tilde{w}_k$. For both theorems, we choose $\alpha_{\text{down}} = (8\omega_{\text{down}})^{-1}$. We denote $\Phi(\gamma) := (1 + \omega_{\text{up}})(1 + 64\gamma L_{\text{down}}^2)$.

**Limit learning rate:** There exists a maximal learning rate to ensure convergence. More specifically, we define $\gamma_{\max} := \min(\gamma_{\text{up}}^{\max}, \gamma_{\text{dwn}}^{\max}, \gamma_{\max}^{\text{up}})$, where $\gamma_{\text{up}}^{\max} := (2L(1 + \omega_{\text{up}}/N))^{-1}$ corresponds to the classical constraint on the learning rate in the unidirectional regime [see 34, 36], $\gamma_{\text{dwn}}^{\max} := (8L_{\text{down}})^{-1}$ is a similar constraint coming from the downlink compression, and $\gamma_{\max}^{\text{up}} := (8\sqrt{2}L_{\text{down}}\sqrt{8\omega_{\text{down}} + \omega_{\text{up}}/N})^{-1}$ is a combined constraint that arises when controlling the variance term $\Phi$. Overall, these constraints are weaker than in the “degraded” framework [29, 36], in which $\gamma_{\text{Dore}}^{\max} \leq (8L(1 + \omega_{\text{down}})(1 + \omega_{\text{up}}/N))^{-1}$. Especially, in the regime in which $\omega_{\text{up,down}} \to \infty$ and $\omega_{\text{down}} \approx \omega_{\text{up}} := \omega$, the maximal learning rate for MCM is $(L_{\omega}^2)^{-1}$, 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 $\bar{L}$ such that $\gamma_{\max} = (2\bar{L})^{-1}$.

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

$$V_k \leq (1 - \gamma_k\mu)V_{k-1} - \gamma_k E[F(\bar{w}_{k-1}) - F(w_*)] + \frac{\gamma_k^2\sigma^2\Phi(\gamma_k)}{N}$$

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

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

**Limit Variance (Equation 4).** For a constant $\gamma$, the variance term (i.e., term proportional to $\sigma^2$) in Equation (4) is upper bounded by $\frac{\gamma^2\sigma^2}{\mu K N b}(1 + \omega_{\text{up}})(1 + 64\gamma L_{\text{down}}^2)$. The impact of the downlink compression is attenuated by a factor $\gamma$. As $\gamma$ 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 $\gamma^2\sigma^2(1 + \omega_{\text{up}})(1 + \omega_{\text{down}})/N$: (1) for Dore, see Corollary 1 in Liu et al. [29] (who indicates $(1 - \rho)^{-1} \geq (1 + \omega_{\text{up}}/N)(1 + \omega_{\text{down}}))$, (2) for Artemis see Table 2 and Th. 3 point 2 in [36], (3) for [14], see Theorem I.1. (with $\gamma D_i^{\prime} \propto \gamma^2\sigma^2(1 + \omega_{\text{up}})(1 + \omega_{\text{down}})/N$).

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

\[\text{The dependency in } \omega^{3/2} \text{ is similar to the one obtained by Horváth et al. [18] in unidirectional compression in the non-convex case (Theorem 4).}\]
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 \( \omega_{\text{down}} \) 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 \( \epsilon \) expected error) is thus at first order \( O_{\epsilon \to 0}(\frac{\sigma^2(1+\omega_{\text{up}})}{\mu N\epsilon}) \). Again, this matches the complexity of Diana [18, see Theorem 1 and Corollary 1] and is smaller by a factor \( 1 + \omega_{\text{down}} \) than the one of Artemis, Dore, DIA\( \text{A}\)\( \text{S}\)\( \text{er}\)–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 \( \gamma \leq \gamma_{\text{max}} \), we have, for \( \bar{w}_k = \frac{1}{K} \sum_{i=0}^{k-1} w_i \),

\[
\gamma \mathbb{E}[F(w_{k-1}) - F(w_*)] \leq V_{k-1} - V_k + \frac{\gamma^2 \sigma^2 \Phi(\gamma)}{N\epsilon} \implies \mathbb{E}[F(\bar{w}_k) - F_*] \leq \frac{V_0}{\gamma_k} + \frac{\gamma^2 \Phi(\gamma)}{N\epsilon}.
\]

(6)

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

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

(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 \( \frac{(1 + \omega_{\text{up}})}{N\epsilon} \) independently of the downlink compression. Again, for previous algorithms with double compression the complexity is at least \( O (\frac{(1 + \omega_{\text{up}})}{N\epsilon}) \) (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 \( \|\bar{w}_k - w_k\|^2 / N\epsilon \) (which is upper-bounded by \( \omega_{\text{down}} \mathbb{E}[\|\mathcal{Y}_k\|^2 / N\epsilon] \)) and to build the Lyapunov function defined in Theorems 1 and 2.

This result highlights the impact of the downlink memory term. Without memory, i.e., with \( \alpha_{\text{down}} = 0 \), the variance of the local model \( \|\bar{w}_k - w_k\|^2 \) increases with the number of iterations. On the other hand, if \( \alpha_{\text{down}} \) 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 \( \gamma \leq (8\omega_{\text{down}} L)^{-1} \) and \( \alpha \leq (4\omega_{\text{down}})^{-1} \), then for all \( k \) in \( \mathbb{N} \):

\[
\mathbb{E}[\mathcal{Y}_k] \leq \left( 1 - \frac{\omega_{\text{down}}}{2} \right) \mathbb{E}[\mathcal{Y}_{k-1}] + 2\gamma^2 \left( \frac{1}{\alpha_{\text{down}}} + \omega_{\text{up}}\right) N \mathbb{E} [\|\nabla F(\bar{w}_{k-1})\|^2] + \frac{2\gamma^2 \sigma^2 (1 + \omega_{\text{up}})}{N\epsilon}.
\]

This bound provides a recursive control on \( \mathcal{Y}_k \). Beyond the \( (1 - \alpha_{\text{down}}) \) 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 \( \omega_{\text{up}}, \omega_{\text{down}}, \) 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.
1. Worker dependent compression.

A first simple situation is the case in which workers are allowed As mentioned in the introduction, our analysis of perturbed iterate in the context of double com-

Appendix E. Th. S11 in Appendix E.4 ensures convergence for a non-convex

This is described in the description of the adaptation to partial participation in [36], in the remark

Indeed it depends on many factors like the selected operators of compression, the upload/downlink

4.1 Communication and convergence trade-offs

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.

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

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

We further define \(\Xi_k = \frac{1}{N} \sum_{i=1}^{N} \|h^i_k - \nabla F_i(w_*)\|^2\), 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). \(\text{How to split a given communication budget between uplink and downlink to optimize the convergence is an open question which is intrinsically related to the situation. Indeed it depends on many factors like the selected operators of compression, the upload/downlink speed or the number of participating workers at each iteration. However, our approach provides some insights on this question. Because asymptotically the impact of double compression is marginal, for a fixed budget, Theorem 2 suggests to strongly compress on the downlink direction (which leads to a large } \omega_{dwn}, \text{ but to perform a weaker compression in the uplink direction).}\)

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

<table>
<thead>
<tr>
<th>Problem</th>
<th>Diana</th>
<th>Artemis, Dore</th>
<th>MCM, Rand-MCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>(L_{\gamma_{\text{max}}} \propto)</td>
<td>(1/(1 + \omega_{\text{up}}))</td>
<td>(1/(1 + \omega_{\text{up}})[1 + \omega_{\text{down}}])</td>
<td>(1/(1 + \omega_{\text{down}})[\sqrt{1 + \omega_{\text{up}}} \wedge 1/(1 + \omega_{\text{up}})])</td>
</tr>
<tr>
<td>Lim. var. (\propto \gamma^2 \sigma^2 / n \times)</td>
<td>(1/(1 + \omega_{\text{up}}))</td>
<td>(1 + \omega_{\text{down}}[1 + \omega_{\text{down}}])</td>
<td>(1/(1 + \omega_{\text{down}})[1 + \omega_{\text{down}}])</td>
</tr>
<tr>
<td>Convex</td>
<td>Rate on init. cond. (SC)</td>
<td>((1 - \gamma)^k)</td>
<td>((1 - \gamma)^k)</td>
</tr>
<tr>
<td></td>
<td>Complexity</td>
<td>((1 + \omega_{\text{up}})/\mu N)</td>
<td>((1 + \omega_{\text{down}})[1 + \omega_{\text{down}}])\mu N)</td>
</tr>
<tr>
<td></td>
<td>Complexity</td>
<td>((\omega_{\text{up}} + 1)/\epsilon^2)</td>
<td>((1 + \omega_{\text{down}})/(1 + \omega_{\text{down}}))\epsilon^2)</td>
</tr>
</tbody>
</table>
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 $G_q$, $q \in [G]$, all workers share the same memory $(H_k^q)$ and receive the same update $C_{dwn, q}(w_{k+1} - H_k^q)$. 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 $\omega_{dwn}$. Keeping in mind that the dominating term in the rate is independent of $\omega_{dwn}$, 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 $\omega_{up}$ (that is, the effect would be the most significant if $\omega_{up}$ 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 $E[||\nabla F(\bar{w}_{k-1}) - E[\nabla F(\bar{w}_{k-1})]||^2 \mid w_{k-1}]$.

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

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

with $\Phi_{Rd}(\gamma) = (1 + \omega_{up}) \left(1 + \frac{4 \gamma^2 L^2 \omega_{dwn} (1 + \frac{\omega_{up}}{N})}{C} \right)$ and $C = N$ for Rand-MCM, $C = G$ Rand-MCM-G, and $C = 1$ 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 $\Phi$ is multiplied by an additional $\gamma \left( \frac{C}{E} + \frac{\omega_{up}}{N} \right)$ (w.r.t. Theorem 2), which is decaying at the same rate as $\gamma$. Besides, the proof for Rand-MCM is substantially modified, as $E[\nabla F(\bar{w}_{k-1})]$ 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 $\omega_{up}$ 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.
In these experiments, we provide results on the log of the excess loss $F(w_k) - F_\star$, 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 $\log_{10}(F(w_k) - F_\star)$. 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 illustrate that when $\sigma^2 = 0$, we recover a linear convergence\(^3\), 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 phishing, 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.

![Figure 2: Convergence on neural networks.](image)

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

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

**Saturation level.** All experiments are performed with a constant learning rate $\gamma$ 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 $\sigma^2$. Theorem 2 states that the variance of MCM is proportional to $\omega_{up}$, this is experimentally observed on Tables 3 and 4 and Fig. 2 and 3: MCM meets Diana while Artemis and Dore saturate at a higher level (scaling as $\omega_{up} \times \omega_{down}$). These trade-offs are preserved with optimized learning rates.

\(^3\)Even stronger, we show in experiments that we recover a linear rate if we have $\sigma_\star = 0$ (the noise over stochastic gradient computation at the optimum point $w_\star$).
The impact of randomization is noticeable on randomized smoothing. Figure 2b 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.

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

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


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 15
  A.1 Compression Operators .......................... 15
  A.2 Relation to Randomized Smoothing ............. 16
B Experiments 16
  B.1 Convex settings ................................. 17
  B.2 Experiments in deep learning .................. 23
  B.3 Wall clock time ................................ 23
  B.4 Hardware and Carbon footprint ................ 23
C Technical results 25
  C.1 Basic inequalities .............................. 25
  C.2 Two lemmas ................................... 26
D The Ghost algorithm 27
  D.1 Motivation, definition of Ghost and proof sketch .. 27
  D.2 Convergence of Ghost, complete proof .......... 28
E Proofs for MCM (and Rand-MCM) 31
  E.1 Control of the Variance of the local model for MCM (Theorem 3) ........ 31
  E.2 Convex case (Theorem 2) ....................... 34
  E.3 Strongly-convex case (Theorem 1) ............... 36
  E.4 Non-convex case (extra theorem) ............... 39
  E.5 Proof for Rand-MCM (Theorem 4) ............... 42
F Proofs in the quadratic case for MCM and Rand-MCM 42
  F.1 Two other lemmas ................................ 43
  F.2 Control of the Variance of the local model for quadratic function (both MCM and Rand-MCM) .......... 46
  F.3 Proof for quadratic function (Theorem 5) ......... 48
G Adataptation to the heterogeneous scenario 51
  G.1 Control of the uplink memory .................. 51
  G.2 Proofs for MCM ............................... 53
H Neurips Checklist 58

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 operators, following similar lines of proof as [6]. The analysis could be extended to biased uplink operators, following similar lines of proof as stated by Assumption 1, we consider only unbiased operators: for instance $\alpha$-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 $\hat{w}_{k-1}$ knowing $w_{k-1}$ is $w_{k-1}$. The extension for the downlink operator seems more difficult as our analysis relies on numerous occurrences on the fact that the expectation of $\hat{w}_{k-1}$ knowing $w_{k-1}$ is $w_{k-1}$.

### A.2 Relation to Randomized Smoothing

Our approach can also be related to randomized smoothing. Formally, $\nabla F(\tilde{w}_{k-1})$ can be considered as an unbiased gradient of the smoothed function $F_\rho$ at point $w_{k-1}$, with $F_\rho: w \mapsto E[F(w + \tilde{w}_{k-1} - w_{k-1})]$. Then $E(\nabla F_\rho(w_{k-1} - w_{k-1})) = E(\nabla F_\rho(w_{k-1} - w_{k-1}) - w_{k-1} - w_{k-1})$. One key aspect is that the condition number $\mu_\rho/L_\rho$ of $F_\rho$ is always larger (better) than the one for $F$. However, the minimum of $F_\rho$ 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 $\hat{w}_{i,k}$ of the global model $w_k$ kept on central server. Thus, we compute the local gradient at a perturbed point $w_k + \delta_{i,k}$. 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 $\omega_{\text{up}} = \omega_{\text{dwn}}$.

In addition, we choose $\alpha_{\text{up}} = \alpha_{\text{dwn}} = \frac{1}{2(1 + \omega_{\text{up}}/\omega_{\text{dwn}})}$.

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

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

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

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

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

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

\[ \log_{10}(F(w_k) - F(w_*)) \]

(a) X axis in # iterations.

(b) X axis in # bits.
We observe that partial participation leads to an increase of the variance for all algorithms. Furthermore, 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:}

![Graphs](image-url)

**Figure S4:** A9A with \( b = 50, \gamma = 1/L \).

**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.

![Graphs](image-url)

**Figure S6:** X axis in # iterations using random sparsification with \( p = 0.1 \).
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 $\bar{H}_k = \frac{1}{N} \sum_{i=1}^{N} H_i^k$ and each worker $i$ holds there own memory $H_i^k$.

**$N$ memories**

\[
\begin{align*}
\Omega_{k+1}^i &= w_{k+1}^i - H_i^k, \\
\hat{w}_{k+1}^i &= H_i^k + C_{dwn,i}(\Omega_{k+1}^i) \\
H_{k+1}^i &= H_i^k + \alpha_{dwn} C_{dwn,i}(\Omega_{k+1}^i).
\end{align*}
\]

**1 memories**

\[
\begin{align*}
\Omega_{k+1} &= w_{k+1}^1 - \bar{H}_k, \\
\hat{w}_{k+1}^1 &= \bar{H}_k + C_{dwn,1}(\Omega_{k+1}) \\
H_{k+1} &= H_k + \alpha_{dwn} C_{dwn,1}(\Omega_{k+1}) \\
\bar{H}_{k+1} &= \bar{H}_k + \frac{\alpha_{dwn}}{N} \sum_{i=1}^{N} C_{dwn,i}(\Omega_{k+1}).
\end{align*}
\]

(S1)
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:

\[
\forall i \in [1, N], \Delta_i^k = g_{i+1}^k(\hat{w}_k) - h_i^k \\
w_{k+1} = w_k - \frac{1}{N} \sum_{i=1}^{N} C_{\text{up}}(\Delta_i^k) + h_i^k \\
\hat{w}_{k+1} = \hat{w}_k - \gamma C_{\text{dwn}} \left( \frac{1}{N} \sum_{i=1}^{N} C_{\text{up}}(\Delta_i^k) + h_i^k \right) \\
h_{i+1}^k = h_i^k + \alpha_{\text{up}} C_{\text{up}}(\Delta_i^k).
\]

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 \( H_{k+1} = \hat{w}_{k+1} \). Indeed by definition we have \( H_{k+1} = H_k + \alpha \hat{\Omega}_{k+1} \), and furthermore, when we rebuild the compressed model on remote device, we have: \( \hat{w}_{k+1} = \hat{\Omega}_{k+1} + H_k \). **In this case, we use the compressed model as memory.**
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.

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

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 $\alpha$ 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 $\alpha$

On Figure S12, we plot the value of the excess loss obtained after 250 epochs w.r.t. to the value of $1/\alpha (\omega_{dwn} + 1)$. We observe that if $\alpha$ is too big, MCM converges slowly; but after reaching a threshold, the value of $\alpha$ does not impact anymore the rate of convergence. This confirms theory that suggests to use the largest possible $\alpha_{dwn}$ but smaller than a given value. The condition $\alpha_{dwn} \leq 1/4(\omega_{dwn} + 1)$ 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 $\alpha_{dwn} = 1/2(\omega_{dwn} + 1)$ as in [36] (see condition S19 in Theorem S7), and this choice is confirmed by Figure S12.
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.
Table S3: Wall clock time on FE-MNIST with $b = 128$ and $s = 2^2$.

<table>
<thead>
<tr>
<th>Compression regime</th>
<th>Computation time for 150 epoch</th>
</tr>
</thead>
<tbody>
<tr>
<td>No compression (SGD)</td>
<td>15421s</td>
</tr>
<tr>
<td>Compression on uplink</td>
<td>16773s, ratio: 1.08</td>
</tr>
<tr>
<td>Compression on uplink and downlink</td>
<td>16769s, ratio: 1.08</td>
</tr>
</tbody>
</table>

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 1,270km 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 $\mathbb{N}^*$, $\hat{g}_k = \frac{1}{N} \sum_{i=1}^{N} \hat{g}_i^k(w_{k-1})$. 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:

\[
\langle a, b \rangle \leq \frac{\|a\|^2}{2\beta} + \frac{\beta \|b\|^2}{2},
\]

\[
\|a + b\|^2 \leq (1 + \frac{1}{\beta}) \|a\|^2 + (1 + \beta) \|a\|^2,
\]

\[
\|a + b\|^2 \leq 2 \left(\|a\|^2 + \|b\|^2\right),
\]

\[
|\langle a, b \rangle| \leq \|a\| \cdot \|b\| \quad \text{(Cauchy-Schwarz inequality)},
\]

\[
\langle a, b \rangle \leq \frac{1}{2} \left(\|a\|^2 + \|b\|^2 - \|a - b\|^2\right) \quad \text{(Polarization identity)}.
\]

Below, we recall Jensen’s inequality.

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

\[
f(E(X)) \leq E(f(X)).
\]

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

**Lemma S1.** Let a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with $\Omega$ a sample space, $\mathcal{A}$ an event space, $\mathbb{P}$ 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:

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

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

\[
E\left[\|X - E[X | \mathcal{F}]\|^2 \mid \mathcal{F}\right] \leq E\left[\|X - Y\|^2 \mid \mathcal{F}\right].
\]

**Assumption 5 (Cocoercivity).** We suppose that for all $k$ in $\mathbb{N}$, stochastic gradients functions $(g_i^k)_{i \in [1, N]}$ 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:

\[
E[\|g_i^k(w_1) - g_i^k(w_2)\|^2] \leq L \langle \nabla F_i(w_1) - \nabla F_i(w_2), w_1 - w_2 \rangle.
\]

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 $\mathbb{R}^d$:

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

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

\[
\langle \nabla F(x), w - w_* \rangle \geq \frac{1}{2} \|F(w) - F(w_*)\| + \frac{1}{2} \left(\mu \|w - w_*\|^2 + \frac{1}{L} \|\nabla F(w)\|^2\right).
\]
C.2 Two lemmas

In this subsection, we give two lemmas required to prove the convergences of Ghost, 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:

\[
E[w_k - H_{k-1} | w_{k-1}] = (1 - \alpha_{dwn})(w_{k-1} - H_{k-2}) - \epsilon E[\nabla F(\hat{w}_{k-1}) | w_{k-1}].
\]

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

\[
E[w_k - H_{k-1} | w_{k-1}] = E[w_{k-1} - \gamma \hat{g}_k(\hat{w}_{k-1}) - (H_{k-2} + \alpha_{dwn} C(w_{k-1} - H_{k-2})) | w_{k-1}]
= (w_{k-1} - H_{k-2}) - \alpha_{dwn} E[C(w_{k-1} - H_{k-2}) | w_{k-1}] - \gamma E[\hat{g}_k | w_{k-1}],
\]

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 \(\sigma^2\).

**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.:

\[
E[\|\tilde{g}_k\|^2 | \hat{w}_{k-1}] \leq \left(1 + \frac{\omega_{up}}{N^b}\right) \|\nabla F(\hat{w}_{k-1})\|^2 + \frac{\sigma^2(1 + \omega_{up})}{N^b},
\]

\[
E[\tilde{g}_k - \nabla F(\hat{w}_{k-1})]^2 | \hat{w}_{k-1}] \leq \frac{\omega_{up} N \|\nabla F(\hat{w}_{k-1})\|^2 + \sigma^2(1 + \omega_{up})}{N^b}.
\]

**Interpretation:**

- If \(\omega_{up} = 0\) (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:

\[
E[\|C(\nabla F(w_{k-1})) - \nabla F(w_{k-1})\|^2 \leq \omega_{up} \|\nabla F(w_{k-1})\|^2\] which is consistent with Assumption 1.

**Proof.** Let \(k\) in \(\mathbb{N}^*\), then

\[
E[\|\tilde{g}_k\|^2 | \hat{w}_{k-1}] = \|\nabla F(\hat{w}_{k-1})\|^2 + E[\|\tilde{g}_k - \nabla F(\hat{w}_{k-1})\|^2 | \hat{w}_{k-1}].
\]

Secondly:

\[
E[\|\tilde{g}_k - \nabla F(\hat{w}_{k-1})\|^2 | \hat{w}_{k-1}]
= E\left[\left\| \frac{1}{N} \sum_{i=1}^{N} \left( \hat{g}_k(\hat{w}_{k-1}) - \nabla F(\hat{w}_{k-1}) \right) \right\|^2 | \hat{w}_{k-1}\right]
= E\left[\left\| \frac{1}{N} \sum_{i=1}^{N} \left( \hat{g}_k(\hat{w}_{k-1}) - g_k(\hat{w}_{k-1}) + g_k(\hat{w}_{k-1}) - \nabla F(\hat{w}_{k-1}) \right) \right\|^2 | \hat{w}_{k-1}\right]
= E\left[\left\| \frac{1}{N} \sum_{i=1}^{N} \left( \hat{g}_k(\hat{w}_{k-1}) - g_k(\hat{w}_{k-1}) \right) \right\|^2 | \hat{w}_{k-1}\right]
+ E\left[\left\| \frac{1}{N} \sum_{i=1}^{N} \left( g_k(\hat{w}_{k-1}) - \nabla F(\hat{w}_{k-1}) \right) \right\|^2 | \hat{w}_{k-1}\right],
\]

\(\text{Ghost is defined in Appendix D.1.}\)
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:

\[ E \left[ \| \tilde{g}_k - \nabla F(\tilde{w}_{k-1}) \|^2 \mid \tilde{w}_{k-1} \right] = \frac{1}{N^2} \sum_{i=1}^{N} E \left[ \left\| g^i_k(\tilde{w}_{k-1}) - g^i_k(\tilde{w}_{k-1}) \right\|^2 \mid \tilde{w}_{k-1} \right] + \frac{1}{N^2} \sum_{i=1}^{N} E \left[ \| g^i_k(\tilde{w}_{k-1}) - \nabla F(\tilde{w}_{k-1}) \|^2 \mid \tilde{w}_{k-1} \right]. \]

Then, for any \( i \) in \([1, N]\), as \( E \left[ \left\| g^i_k(\tilde{w}_{k-1}) - g^i_k(\tilde{w}_{k-1}) \right\|^2 \mid \tilde{w}_{k-1} \right] = E \left[ \left\| g^i_k(\tilde{w}_{k-1}) \right\|^2 \mid \tilde{w}_{k-1} \right] \), and using Assumption 1 we have:

\[ E \left[ \left\| g^i_k(\tilde{w}_{k-1}) \right\|^2 \mid \tilde{w}_{k-1} \right] = \frac{\omega_{up}}{N^2} \sum_{i=1}^{N} E \left[ \| g^i_k(\tilde{w}_{k-1}) \|^2 \mid \tilde{w}_{k-1} \right] + \frac{1}{N^2} \sum_{i=1}^{N} E \left[ \| g^i_k(\tilde{w}_{k-1}) - \nabla F(\tilde{w}_{k-1}) \|^2 \mid \tilde{w}_{k-1} \right]. \]

Furthermore

\[ E \left[ \left\| g^i_k(\tilde{w}_{k-1}) \right\|^2 \mid \tilde{w}_{k-1} \right] = E \left[ \left\| g^i_k(\tilde{w}_{k-1}) - \nabla F(\tilde{w}_{k-1}) \right\|^2 \mid \tilde{w}_{k-1} \right] + \frac{\omega_{up}}{N} \left( \| \nabla F(\tilde{w}_{k-1}) \|^2 + \frac{\sigma^2(1 + \omega_{up})}{Nb} \right), \]

from which we derive the two inequalities of the lemma.

\[ \square \]

### 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:

\[ w_{k+1} = w_k - \gamma \frac{1}{N} \sum_{i=1}^{N} g^i_{k+1}(\tilde{w}_k) \quad \text{and} \quad \tilde{w}_{k+1} = w_k - \gamma \tilde{C}_{\text{dwn}} \left( \frac{1}{N} \sum_{i=1}^{N} g^i_{k+1}(\tilde{w}_k) \right). \]  \( \text{(S10)} \)

While the global model is unchanged (1st line), the local model \( \tilde{w}_{k+1} \) (2nd 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 \( \nabla F(w_{-1}) = 0 \):

\[ E \left[ \left\| w_k - \tilde{w}_k \right\|^2 \mid \tilde{w}_{k-1} \right] \leq \gamma^2 \omega_{\text{dwn}} \left( 1 + \frac{\omega_{up}}{N} \right) \| \nabla F(\tilde{w}_{k-1}) \|^2 + \frac{\gamma^2 \omega_{\text{dwn}}(1 + \omega_{up}) \sigma^2}{Nb}. \]
Proof. The proof of Proposition S2 is straightforward using Definition 1. Let k in \( \mathbb{N} \), by Definition 1 we have:

\[
\| \hat{w}_k - \hat{w}_{k-1} \|^2 = \left\| \left( \hat{w}_{k-1} - \frac{1}{N} \sum_{i=1}^{N} \hat{g}_k^i(\hat{w}_{k-1}) \right) - \left( \hat{w}_{k-1} - \frac{1}{N} \sum_{i=1}^{N} \hat{g}_k^i(\hat{w}_{k-1}) \right) \right\|^2.
\]

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

\[
E \left[ \| \hat{w}_k - \hat{w}_{k-1} \|^2 \right] = \gamma^2 \omega_{dwn} E \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \hat{g}_k^i(\hat{w}_{k-1}) \right\|^2 \right] = \gamma^2 \omega_{dwn} \| \hat{g}_k \|^2,
\]

and Lemma S3 gives the upper bound \( E \left[ \| \hat{g}_k \|^2 \right] \hat{w}_{k-1} \).

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 \( \nabla F(\hat{w}_{k-1}) \). 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 \( E \| w_k - w_* \|^2 \) as \( E \| w_k - w_* \|^2 - 2\gamma E \langle \nabla F(w_k - 1), w_k - w_* \rangle + \gamma^2 E \| \hat{g}_k(\hat{w}_{k-1}) \|^2 \). 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 \( E [ \hat{w}_{k-1} | w_{k-1} = w_{k-1} ] \), we further decompose it as

\[
-2\gamma \langle \nabla F(\hat{w}_{k-1}), \hat{w}_{k-1} - w_* \rangle + 2\gamma E \langle \nabla F(\hat{w}_{k-1}) - \nabla F(w_{k-1}), w_{k-1} - \hat{w}_{k-1} \rangle.
\]

The first part of eq. (S11), corresponds to a “strong contraction”: by (strong-)convexity, we can upper bound it by \( -2\gamma (\mu \| \hat{w}_{k-1} - w_* \|^2 + F(\hat{w}_{k-1}) - F_* \) 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 \( -2\gamma \| \nabla F(\hat{w}_{k-1}) \|^2 / L \). This is a crucial term: we “gain” something of the order of a squared norm of the gradient at \( \hat{w}_{k-1} \), 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 \( \gamma L (1 + \omega_{up}/N) \leq \frac{1}{2} \),

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

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 \( 2\gamma^3 \omega_{dwn} L (1 + \omega_{up}/N) \leq \gamma/(2L) \), then these terms cancel out. This is equivalent to \( 2\gamma L \sqrt{\omega_{dwn} (1 + \omega_{up}/N)} \leq 1 \). It is natural to chose \( \gamma \leq 1/(2L \max(1 + \omega_{up}/N, 1 + \omega_{dwn})) \).

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.
Taking expectation conditionally to \(\omega\) and thus:

\[
E \left[ \|w_k - w_*\|^2 \right] \leq \|w_k - v_{k-1}\|^2 - 2\gamma \langle \tilde{g}_k, w_k - w_* \rangle + \gamma^2 \|g_k\|^2.
\]

Next, we expend the inner product as following:

\[
\|w_k - w_*\|^2 \leq \|w_k - v_{k-1}\|^2 - 2\gamma \langle \tilde{g}_k, \tilde{w}_{k-1} - w_* \rangle - 2\gamma \langle \tilde{g}_k, w_k - \tilde{w}_{k-1} \rangle + \gamma^2 \|g_k\|^2.
\]

Taking expectation conditionally to \(w_{k-1}\), and using \(E[\tilde{g}_k | w_{k-1}] = E[E[\tilde{g}_k | \tilde{w}_{k-1}] | w_{k-1}] = E[\nabla F(\tilde{w}_{k-1}) | w_{k-1}]\), we obtain:

\[
E \left[ \|w_k - w_*\|^2 \right] \leq \|w_k - v_{k-1}\|^2 - 2\gamma E \left[ \langle \nabla F(\tilde{w}_{k-1}), \tilde{w}_{k-1} - w_* \rangle \right] | w_{k-1} | - 2\gamma E \left[ \langle \nabla F(\tilde{w}_{k-1}), w_k - \tilde{w}_{k-1} \rangle \right] | w_{k-1} | + \gamma^2 E \left[ \|g_k\|^2 \right] | w_{k-1} |.
\]

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

\[
E \left[ \|w_k - w_*\|^2 \right] | w_{k-1} | \leq \|w_k - v_{k-1}\|^2 - 2\gamma E \left[ \langle \nabla F(\tilde{w}_{k-1}), \tilde{w}_{k-1} - w_* \rangle \right] | w_{k-1} | - 2\gamma E \left[ \langle \nabla F(\tilde{w}_{k-1}) - \nabla F(w_{k-1}), w_k - \tilde{w}_{k-1} \rangle \right] | w_{k-1} | + \gamma^2 \left( 1 + \frac{\omega_{up}}{Nb} \right) E \left[ \|\nabla F(\tilde{w}_{k-1})\|^2 \right] | w_{k-1} | + \frac{\sigma^2 (1 + \omega_{up})}{Nb}.
\]

In the upper inequality:

1. the term \(E \left[ \langle \nabla F(\tilde{w}_{k-1}), \tilde{w}_{k-1} - w_* \rangle \right] | w_{k-1} |\) allows the “strong contraction”

2. the terms \(E \left[ \|\nabla F(\tilde{w}_{k-1}) - \nabla F(w_{k-1}), w_k - \tilde{w}_{k-1} \rangle \right] | w_{k-1} |\) are two positives terms that we treat as residuals.

3. the last term \(\sigma^2 (1 + \omega_{up})/(Nb)\) is due to the stochastic noise.

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

\[
- \gamma E \left[ \|w_k - v_{k-1}\|^2 \right] | w_{k-1} | \leq 2\gamma E \left[ \langle \nabla F(\tilde{w}_{k-1}) - \nabla F(w_{k-1}), \tilde{w}_{k-1} - w_{k-1} \rangle \right] | w_{k-1} | \leq 2\gamma L E \left[ \|w_{k-1} - w_{k-1}\|^2 \right] | w_{k-1} |,
\]

and thus:

\[
E \left[ \|w_k - w_*\|^2 \right] | w_{k-1} | \leq \|w_k - v_{k-1}\|^2 - 2\gamma E \left[ \langle \nabla F(\tilde{w}_{k-1}), \tilde{w}_{k-1} - w_* \rangle \right] | w_{k-1} | + 2\gamma L E \left[ \|w_{k-1} - w_{k-1}\|^2 \right] | w_{k-1} | + \gamma^2 \left( 1 + \frac{\omega_{up}}{N} \right) E \left[ \|\nabla F(\tilde{w}_{k-1})\|^2 \right] | w_{k-1} | + \frac{\sigma^2 (1 + \omega_{up})}{Nb}.
\]
Now, using convexity with Proposition S1:
\[
E \left[ \| w_k - w_* \|^2 \Bigg| w_{k-1} \right] \leq \| w_{k-1} - w_* \|^2
\]
\[
- \gamma E \left[ \left( F(\hat{w}_{k-1}) - F(w_*) + \frac{1}{L} \| \nabla F(\hat{w}_{k-1}) \|^2 \right) \Bigg| \hat{w}_{k-1} \right]
\]
\[
+ 2\gamma L E \left[ \| \hat{w}_{k-1} - w_{k-1} \|^2 \bigg| w_{k-1} \right]
\]
\[
+ \gamma^2 \left( 1 + \frac{\omega_{up}}{N} \right) E \left[ \| \nabla F(\hat{w}_{k-1}) \|^2 \bigg| w_{k-1} \right] + \frac{\gamma^2 \sigma^2 (1 + \omega_{up})}{N}. \tag{S14}
\]
Taking the full expectation (without conditioning over any random vectors), and because invoking Jensen inequality (S7) leads to \( E[F(\hat{w}_{k-1})] \geq E[F(w_{k-1})] \), we finally obtain this intermediate result:
\[
E \left[ \| w_k - w_* \|^2 \right] \leq E \left[ \| w_{k-1} - w_* \|^2 \right] - \gamma \left( E \left[ F(w_{k-1}) \right] - F(w_*) \right)
- \frac{\gamma}{2L} E \left[ \| \nabla F(\hat{w}_{k-1}) \|^2 \right]
+ 2\gamma L E \left[ \| \hat{w}_{k-1} - w_{k-1} \|^2 \right] + \frac{\gamma^2 \sigma^2 (1 + \omega_{up})}{N}. \tag{S14}
\]
where we considered that \( \gamma L (1 + \omega_{up}/N) \leq 1/2 \), which implies that \( \gamma \left( 1 - \gamma L (1 + \omega_{up}/N) \right) \geq \frac{\gamma}{2} \).

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:
\[
E \left[ \| w_k - \hat{w}_k \|^2 \bigg| \hat{w}_{k-1} \right] \leq \gamma^2 \omega_{dwn} \left( 1 + \frac{\omega_{up}}{N} \right) \| \nabla F(\hat{w}_{k-1}) \|^2 + \frac{\gamma^2 \omega_{dwn}(1 + \omega_{up})\sigma^2}{N}. \tag{S15}
\]
Defining \( V_k := E \left[ w_k - w_* \right] + \frac{\gamma}{2L} E \left[ \| \nabla F(\hat{w}_{k-1}) \|^2 \right] + C \left[ \| \hat{w}_k - w_k \|^2 \right] \) with \( C = 2\gamma L \), and combining this two equations as following (S14) + (S15) leads to:
\[
E \left[ \| w_k - w_* \|^2 \right] + C \left[ \| \hat{w}_k - w_k \|^2 \right] + \frac{\gamma}{2L} E \left[ \| \nabla F(\hat{w}_{k-1}) \|^2 \right]
\leq E \left[ \| w_{k-1} - w_* \|^2 \right] - \gamma \left( E \left[ F(w_{k-1}) \right] - F(w_*) \right)
+ 2\gamma L E \left[ \| \hat{w}_{k-1} - w_{k-1} \|^2 \right] + \gamma^2 \sigma^2 \left( 1 + \omega_{up} \right)
+ 2\gamma L \times \gamma^2 \omega_{dwn} \left( 1 + \frac{\omega_{up}}{N} \right) \| \nabla F(\hat{w}_{k-1}) \|^2 + 2\gamma L \times \frac{\gamma^2 \omega_{dwn}(1 + \omega_{up})\sigma^2}{N}.
\]
To ensure a contraction of the Lyapunov function we require:
\[
\gamma^2 \omega_{dwn} \left( 1 + \frac{\omega_{up}}{N} \right) \leq \frac{\gamma}{2L} \iff \gamma L \leq \frac{1}{2\sqrt{\omega_{dwn} \left( 1 + \frac{\omega_{up}}{N} \right)}}.
\]
Under this condition, we obtain:
\[
V_k \leq V_{k-1} - \gamma E \left[ F(w_{k-1}) - F(w_*) \right] + \frac{\gamma^2 \sigma^2 \Phi^\gamma(\gamma)}{N},
\]
where \( \Phi^\gamma(\gamma) := (1 + \omega_{up})(1 + 2\gamma L \omega_{dwn}). \)

By recurrence and for \( k = K \):
\[
V_K \leq V_0 - \sum_{k=1}^{K} \gamma E \left[ F(w_{k-1}) - F(w_*) \right] + \sum_{k=1}^{K} \frac{\gamma^2 \sigma^2 \Phi^\gamma(\gamma)}{N},
\]

30
We denote for \( \gamma \) we define which leads to:

\[
\frac{1}{K} \sum_{k=1}^{K} \mathbb{E} [F(w_{k-1}) - F(w_*)] \leq \frac{V_0 - V_k}{\gamma K} + \frac{\gamma \sigma^2 \Phi^2(\gamma)}{N b}.
\]

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

\[
\frac{\gamma}{K} \sum_{t=1}^{K} \mathbb{E} [F(w_t) - F(w_*)] \leq \frac{\|w_0 - w_*\|^2}{K} + \frac{\gamma \sigma^2 \Phi^2(\gamma)}{N b}.
\]

Note that the bound of \( \gamma L \) encompass the case \( \omega_{dwn} = 0 \) (i.e. no downlink compression), but in the general case of bidirectional compression, we nearly always have \( \omega_{dwn} > 1 \), and thus the dominant term is in fact

\[
\frac{1}{2 \sqrt{\omega_{dwn} \left( 1 + \frac{\omega_{up}}{N} \right)}}.
\]

And by Jensen, it implies that:

\[
\mathbb{E} [F(\bar{w}_K) - F(w_*)] \leq \frac{\|w_0 - w_*\|^2}{\gamma K} + \frac{\gamma \sigma^2 \Phi^2(\gamma)}{N b} \quad \text{with} \quad \Phi^2(\gamma) := (1 + \omega_{up})(1 + 2\omega_{dwn}\gamma L).
\]

\[\square\]

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

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

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

#### 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 \( \alpha_{dwn} \). 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:

\[
\|w_k - H_{k-1}\|^2 = \text{Bias}^2 + 2\gamma^2 (\text{Var}_1 + \text{Var}_2) + 2\alpha_{dwn}^2 \text{Var}_2.
\]  
(S16)

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

\[
\mathbb{E}[\Upsilon_k] \leq \left( 1 - \frac{\alpha_{dwn}}{2} \right) \mathbb{E}[\Upsilon_{k-1}] + 2\gamma^2 \left( \frac{1}{\alpha_{dwn}} + \frac{\omega_{up}}{N} \right) \mathbb{E} \left[ \|\nabla F(\bar{w}_{k-1})\|^2 \right] + \frac{2\gamma^2 \sigma^2 (1 + \omega_{up})}{N b}.
\]
Proof. Let $k$ in $\mathbb{N}$, we recall that by definition:

\[
\begin{aligned}
\Omega_k &= w_k - H_{k-1} \\
\tilde{\Omega}_k &= C_{\text{dwn}}(\Omega_k) \\
\tilde{w}_k &= \tilde{\Omega}_k + H_{k-1} .
\end{aligned}
\]

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

\[
\mathbb{E} \left[ \| w_k - \tilde{w}_k \|^2 \bigg| w_k \right] = \mathbb{E} \left[ \| \tilde{\Omega}_k - \Omega_k \|^2 \bigg| w_k \right] \leq \omega_{\text{dwn}} \| \Omega_k \|^2 .
\]

Next, we perform a bias-variance decomposition:

\[
\|\Omega_k\|^2 = \| w_k - H_{k-1} \|^2 = \| w_k - H_{k-1} - \mathbb{E} [w_k - H_{k-1} \big| w_{k-1}] \|^2
\]
\[+ \| \mathbb{E} [w_k - H_{k-1} \big| w_{k-1}] \|^2
\]
\[+ 2 \langle w_k - H_{k-1} - \mathbb{E} [w_k - H_{k-1} \big| w_{k-1}], \mathbb{E} [w_k - H_{k-1} \big| w_{k-1}] \rangle ,
\]

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

\[
\mathbb{E} [\Upsilon_k \big| w_{k-1}] = \mathbb{E} \left[ \| w_k - H_{k-1} - \mathbb{E} [w_k - H_{k-1} \big| w_{k-1}] \|^2 \bigg| w_{k-1} \right]
\]
\[+ \| \mathbb{E} [w_k - H_{k-1} \big| w_{k-1}] \|^2 .
\]

The first term is the variance $\text{Var}$, and the second term corresponds to the squared bias $\text{Bias}^2$.

Let’s handle first the variance, by definition:

\[
\text{Var} = \mathbb{E} \left[ \| w_k - H_{k-1} - \mathbb{E} [w_k - H_{k-1} \big| w_{k-1}] \|^2 \bigg| w_{k-1} \right]
\]
\[= \mathbb{E} \left[ \| w_{k-1} - \gamma \tilde{g}_k - H_{k-2} - \alpha_{\text{dwn}} C (w_{k-1} - H_{k-2}) \right.
\]
\[\left. - w_{k-1} - \gamma \tilde{g}_k \big| w_{k-1} \bigg| - H_{k-2} - \alpha_{\text{dwn}} \mathbb{E} [C (w_{k-1} - H_{k-2} \big| w_{k-1})] \right]^2 \bigg| w_{k-1} \right] .
\]

After simplification and using eq. (S4):

\[
\text{Var} = \mathbb{E} \left[ \| - \gamma (\tilde{g}_k + \mathbb{E} [\nabla F (\tilde{w}_{k-1}) \big| w_{k-1}]) + \alpha_{\text{dwn}} (C (w_{k-1} - H_{k-2}) \right.
\]
\[\left. - (w_{k-1} - H_{k-2}) \right]^2 \bigg| w_{k-1} \right] \]
\[\leq 2\gamma^2 \mathbb{E} \left[ \| \tilde{g}_k - \mathbb{E} [\nabla F (\tilde{w}_{k-1}) \big| w_{k-1}] \|^2 \bigg| w_{k-1} \right]
\]
\[+ 2\alpha_{\text{dwn}}^2 \mathbb{E} \left[ \| C (w_{k-1} - H_{k-2}) - (w_{k-1} - H_{k-2}) \|^2 \bigg| w_{k-1} \right]
\]
\[\leq 2\gamma^2 \mathbb{E} \left[ \| \tilde{g}_k - \mathbb{E} [\nabla F (\tilde{w}_{k-1}) \big| w_{k-1}] \|^2 \bigg| w_{k-1} \right]
\]
\[+ 2\alpha_{\text{dwn}}^2 \mathbb{E} \left[ \| w_{k-1} \|^2 \bigg| w_{k-1} \right] + 2\alpha_{\text{dwn}}^2 \mathbb{E} \left[ \| w_{k-1} \|^2 \bigg| w_{k-1} \right] .
\]

An interpretation of the above decomposition is that:

• $\text{Var}_1$ is the part of the downlink compression caused by the increment $\tilde{g}_k$, it is similar to Ghost.
• $\text{Var}_2$ is the impact of the propagation of the previous noise.

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

\[
\text{Var}_1 = \mathbb{E} \left[ \| \tilde{g}_k - \mathbb{E} [\nabla F (\tilde{w}_{k-1}) \big| w_{k-1}] \|^2 \bigg| w_{k-1} \right]
\]
\[= \mathbb{E} \left[ \| \tilde{g}_k - \nabla F (\tilde{w}_{k-1}) + \nabla F (\tilde{w}_{k-1}) - \mathbb{E} [\nabla F (\tilde{w}_{k-1}) \big| w_{k-1}] \|^2 \bigg| w_{k-1} \right]
\]
\[= \mathbb{E} \left[ \| \tilde{g}_k - \nabla F (\tilde{w}_{k-1}) \|^2 \bigg| w_{k-1} \right] + \mathbb{E} \left[ \| \nabla F (\tilde{w}_{k-1}) - \mathbb{E} [\nabla F (\tilde{w}_{k-1}) \big| w_{k-1}] \|^2 \bigg| w_{k-1} \right]
\]
\[= \text{Var}_{11} + \text{Var}_{12} ,
\]
the inner product is null given that \( \mathbb{E} [\nabla F(\hat{w}_{k-1}) - \mathbb{E} [\nabla F(\hat{w}_{k-1}) \mid w_{k-1}] \mid w_{k-1}] = 0 \).

Moreover:

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

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

And now we use smoothness for the second term:

\[
\text{Var}_{12} = \mathbb{E}\left[ \| \nabla F(\hat{w}_{k-1}) - \mathbb{E} [\nabla F(\hat{w}_{k-1}) \mid w_{k-1}] \|^2 \mid w_{k-1} \right]
\leq \mathbb{E}\left[ \| \nabla F(\hat{w}_{k-1}) - \nabla F(w_{k-1}) \|^2 \mid w_{k-1} \right]
\leq L^2 \mathbb{E}\left[ \| \hat{w}_{k-1} - w_{k-1} \|^2 \mid w_{k-1} \right]
\leq L^2 \omega_{dwn} \Upsilon_{k-1} \quad \text{using smoothness},
\]

At the end:

\[
\text{Var} \leq 2\gamma^2 \left( \frac{\sigma^2(1 + \omega_{up})}{N b} + \frac{\omega_{up}}{N} \mathbb{E}\left[ \| \nabla F(\hat{w}_{k-1}) \|^2 \mid w_{k-1} \right] + L^2 \omega_{dwn} \Upsilon_{k-1} \right)
+ 2\alpha_{dwn}^2 \omega_{dwn} \Upsilon_{k-1} \tag{S17}
\]

Now we focus on the squared bias \( \text{Bias}^2 \), with Lemma S2:

\[
\text{Bias}^2 = \| \mathbb{E} [w_k - H_{k-1} \mid w_{k-1}] \|^2
= \| (1 - \alpha_{dwn})(w_{k-1} - H_{k-2}) - \gamma \mathbb{E} [\nabla F(\hat{w}_{k-1}) \mid w_{k-1}] \|^2
\leq (1 - \alpha_{dwn})^2 \mathbb{E} [\nabla F(\hat{w}_{k-1}) \mid w_{k-1}] + \gamma^2 (1 + \frac{1}{\alpha_{dwn}}) \mathbb{E} [\nabla F(\hat{w}_{k-1}) \mid w_{k-1}]^2.
\]

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

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

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

\[
\mathbb{E} [\Upsilon_k \mid w_{k-1}] \leq (1 - \alpha_{dwn}) \Upsilon_{k-1} + \gamma^2 \left( 1 + \frac{1}{\alpha_{dwn}} \right) \mathbb{E} [\nabla F(\hat{w}_{k-1}) \mid w_{k-1}]^2
+ 2\gamma^2 \left( \frac{\sigma^2(1 + \omega_{up})}{N b} + \frac{\omega_{up}}{N} \mathbb{E}\left[ \| \nabla F(\hat{w}_{k-1}) \|^2 \mid w_{k-1} \right] \right)
+ 2\gamma^2 \left( L^2 \omega_{dwn} \Upsilon_{k-1} \right)
+ 2\alpha_{dwn}^2 \omega_{dwn} \Upsilon_{k-1}
\leq (1 - \alpha_{dwn} + 2\gamma^2 L^2 \omega_{dwn} + 2\alpha_{dwn} \omega_{dwn} \Upsilon_{k-1}) \| w_{k-1} - H_{k-2} \|^2
+ \gamma^2 \left( 1 + \frac{1}{\alpha_{dwn}} \right) \mathbb{E} [\nabla F(\hat{w}_{k-1}) \mid w_{k-1}]^2
+ 2\alpha_{dwn} \omega_{dwn} \Upsilon_{k-1}
\]

Next, we require:

\[
\begin{cases}
2\alpha_{dwn}^2 \omega_{dwn} \leq \frac{1}{2} \alpha_{dwn} \iff \alpha_{dwn} \leq \frac{1}{8\omega_{dwn}}, \\
2\gamma^2 L^2 \omega_{dwn} \leq \frac{1}{2} \alpha_{dwn} = \frac{1}{2 \gamma^2 \omega_{dwn}}, \text{ by taking } \alpha_{dwn} = \frac{1}{8\omega_{dwn}} \iff \gamma \leq \frac{1}{8\omega_{dwn} L}, \\
1 + \frac{1}{\alpha_{dwn}} \leq \frac{2}{\alpha_{dwn}} \text{ which is not restrictive if } \omega_{dwn} \geq 1.
\end{cases}
\]
Thus, it leads to:

\[
E [Y_k \mid w_{k-1}] \leq \left(1 - \frac{\alpha_{dwn}}{2}\right) Y_{k-1} + \frac{2\gamma^2}{\alpha_{dwn}} E \left[\|\nabla F(\hat{w}_{k-1})\| \mid w_{k-1}\|^2 \right] \\
+ \frac{2\gamma^2 \omega_{up}}{N} E \left[\|\nabla F(\hat{w}_{k-1})\| \mid w_{k-1}\|^2 \right] + \frac{2\gamma^2 \sigma^2 (1 + \omega_{up})}{Nb}.
\]

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

\[
E [Y_k \mid w_{k-1}] \leq \left(1 - \frac{\alpha_{dwn}}{2}\right) Y_{k-1} \\
+ \frac{2\gamma^2}{\alpha_{dwn}} \left(1 + \frac{\omega_{up}}{N}\right) E \left[\|\nabla F(\hat{w}_{k-1})\| \mid w_{k-1}\|^2\right] \\
+ \frac{2\gamma^2 \sigma^2 (1 + \omega_{up})}{Nb}.
\]

Taking the unconditional expectation gives the result.

\[ \square \]

**E.2 Convex case (Theorem 2)**

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

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

\[
\gamma E[F(\bar{w}_{k-1}) - F(w_*)] \leq V_{k-1} - V_k + \frac{\gamma^2 \sigma^2 \Phi(\gamma)}{N b} \implies E[F(\bar{w}_k) - F_*] \leq \frac{V_0}{\gamma k} + \frac{\gamma^2 \Phi(\gamma)}{N b}.
\]

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

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

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

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

\[
E \left[\|w_k - w_*\|^2\right] \leq E \left[\|w_{k-1} - w_*\|^2\right] - \gamma \left(E[F(w_{k-1})] - F(w_*)\right) - \frac{\gamma}{2L} E \left[\|\nabla F(\hat{w}_{k-1})\|^2\right] \\
+ 2\gamma L \omega_{dwn} E \left[\|\hat{w}_{k-1} - w_{k-1}\|^2\right] + \frac{\gamma^2 \sigma^2 (1 + \omega_{up})}{N b},
\]

with Assumption 1, it easily becomes:

\[
E \left[\|w_k - w_*\|^2\right] \leq E \left[\|w_{k-1} - w_*\|^2\right] - \gamma \left(E[F(w_{k-1})] - F(w_*)\right) - \frac{\gamma}{2L} E \left[\|\nabla F(\hat{w}_{k-1})\|^2\right] \\
+ 2\gamma L \omega_{dwn} E \left[Y_{k-1}\right] + \frac{\gamma^2 \sigma^2 (1 + \omega_{up})}{N b}.
\]

**Theorem 3** which is specific to \(\text{MCM}\) gives:

\[
E [Y_k] \leq \left(1 - \frac{\alpha_{dwn}}{2}\right) E [Y_{k-1}] + 2\gamma^2 \left(1 + \alpha_{dwn} + \frac{\omega_{up}}{N}\right) E \left[\|\nabla F(\hat{w}_{k-1})\|\right] + \frac{2\gamma^2 \sigma^2 (1 + \omega_{up})}{N b}.
\]
Defining: $V_k := \mathbb{E} \left[ \|w_k - w_*\|^2 \right] + \gamma LC \mathbb{E} [\Upsilon_k]$ with $C = \frac{4\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}$, and, combining the two last equations:

\[
\mathbb{E} \left[ \|w_k - w_*\|^2 \right] + \gamma LC \mathbb{E} [\Upsilon_k] \leq \mathbb{E} \left[ \|w_{k-1} - w_*\|^2 \right] - \gamma \mathbb{E} [F(w_{k-1}) - F(w_*)] \\
+ 2\gamma L\omega_{\text{dwn}} \mathbb{E} [\Upsilon_k] \\
- \frac{\gamma}{2L} \mathbb{E} \left[ \|\nabla F(\hat{w}_{k-1})\|^2 \right] + \frac{\gamma^2 \sigma^2 (1 + \omega_{\text{up}})}{Nb} \\
+ \left(1 - \frac{\alpha_{\text{dwn}}}{2}\right) \gamma LC \mathbb{E} [\Upsilon_k] \\
+ 2\gamma^3 LC \left(\frac{1}{\alpha_{\text{dwn}}} \frac{\omega_{\text{up}}}{N}\right) \mathbb{E} \left[ \|\nabla F(\hat{w}_{k-1})\|^2 \right] \\
+ 2\gamma^3 L\sigma^2 (1 + \omega_{\text{up}})C 
\]

and reordering the terms gives:

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

We observe that:

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

Secondly, to get the contraction requires

\[
2\gamma^3 LC \left(\frac{1}{\alpha_{\text{dwn}}} \frac{\omega_{\text{up}}}{N}\right) - \frac{\gamma}{2L} \leq 0 \iff \gamma^2 L \leq \frac{1}{4LC \left(\frac{1}{\alpha_{\text{dwn}}} \frac{\omega_{\text{up}}}{N}\right)} \\
\iff \gamma L \leq \frac{1}{4\sqrt{\frac{\omega_{\text{dwn}}}{\alpha_{\text{dwn}}} \left(\frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N}\right)}} ,
\]

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

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

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

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

which implies:

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

35
Finally, by Jensen, for any $K$ in $\mathbb{N}^*$ such that $\gamma L \leq 1$,
\[
\min \left\{ \frac{1}{8\omega_{\text{dwn}}} - \frac{1}{2} (1 + \frac{\omega_{\text{up}}}{N}) ; \frac{1}{4} \left( \frac{\omega_{\text{dwn}}}{\alpha_{\text{dwn}}} \left( \frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N} \right) \right) \right\},
\]
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. \qed

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 $\alpha_{\text{dwn}} \leq \frac{1}{8\omega_{\text{dwn}}}$, for any sequence $(\gamma_k)_{k \geq 0} \leq \gamma_{\text{max}}$, defining $V_k := \mathbb{E}[[w_k - w_*]^2] + 32\gamma L\omega_{\text{dwn}}^2 \mathbb{E}[\Upsilon_k]$, we have:
\[
V_k \leq (1 - \gamma_k \mu) V_{k-1} - \gamma_k \mathbb{E} [F(\bar{w}_{k-1}) - F(w_*)] + \frac{\gamma_k^2 \sigma^2 \Phi(\gamma)}{Nb},
\]
Consequently,
1. if $\sigma^2 = 0$ (noiseless case), for $\gamma_k \equiv \gamma_{\text{max}}$ we recover a linear convergence rate:
\[
\mathbb{E}[[w_K] - w_*]^2] \leq (1 - \gamma_{\text{max}} \mu)^K V_0;
\]
2. if $\sigma^2 > 0$, defining $\bar{L}$ such that $\gamma_{\text{max}} = (2\bar{L})^{-1}$, taking for all $k$ in $\mathbb{N}$, $\gamma_k = 2/(\mu(k + 1) + \bar{L})$, 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 $\lambda_k := \frac{1}{\gamma_{k-1}}$, we have:
\[
\mathbb{E} \left[ F(\bar{w}_K) - F(w_*) \right] \leq \frac{\mu + 2\bar{L}}{4\mu K^2} \|w_0 - w_*\|^2 + \frac{4\sigma^2(1 + \omega_{\text{up}})}{\mu KNb} \left( 1 + \frac{64L\omega_{\text{dwn}}^2}{\mu K N \ln(\mu K + \bar{L})} \right).
\]

**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 $\gamma_k = 2/(\mu(k + 1) + \bar{L})$ that depends on the iteration $k$ in $\mathbb{N}$.
\[
\mathbb{E} \left[ \|w_k - w_*\|^2 \right] \leq \mathbb{E} \left[ \|w_{k-1} - w_*\|^2 \right] - 2\gamma_k \mathbb{E} [\nabla F(\bar{w}_{k-1}) \cdot (\nabla F(\bar{w}_{k-1}) - w_{k-1} - w_*)] + 2\gamma_k L \mathbb{E} \left[ \left\| \bar{w}_{k-1} - w_{k-1} \right\|^2 \right] + \gamma_k^2 \left( 1 + \frac{\omega_{\text{up}}}{N} \right) \mathbb{E} \left[ \left\| \nabla F(\bar{w}_{k-1}) \right\|^2 \right]
\]
\[
+ \frac{\gamma_k^2 \sigma^2 (1 + \omega_{\text{up}})}{Nb}.
\]
Now we apply strong-convexity (eq. (S9) of Proposition S1):
\[
\mathbb{E} \left[ \|w_k - w_*\|^2 \right] \leq \mathbb{E} \left[ \|w_{k-1} - w_*\|^2 \right] + 2\gamma_k L \mathbb{E} \left[ \left\| \bar{w}_{k-1} - w_{k-1} \right\|^2 \right]
\]
\[
- \gamma_k \mathbb{E} [F(\bar{w}_{k-1}) - F(w_*)] - \gamma_k \left( \mu \left\| \bar{w}_{k-1} - w_* \right\|^2 + \frac{1}{\bar{L}} \left\| \nabla F(\bar{w}_{k-1}) \right\|^2 \right)
\]
\[
+ \gamma_k^2 \left( 1 + \frac{\omega_{\text{up}}}{N} \right) \mathbb{E} \left[ \left\| \nabla F(\bar{w}_{k-1}) \right\|^2 \right] + \frac{\gamma_k^2 \sigma^2 (1 + \omega_{\text{up}})}{Nb}.
\]
As $\gamma_k \leq \frac{2}{L} \leq \frac{1}{2L(1 + \omega_{\text{up}}/N)}$, and thus \((1 - \gamma_k L \left(1 + \frac{\omega_{\text{up}}}{N}\right)) \geq 1/2\); this allows to simplify the coefficient of $E \left[\|\nabla F(\tilde{w}_{k-1})\|^2\right]$: 

$$
E \left[\|w_k - w_*\|^2\right] \leq (1 - \gamma_k \mu) \|w_{k-1} - w_*\|^2 - \gamma_k E \left[F(\tilde{w}_{k-1}) - F(w_*)\right] - \frac{\gamma_k}{2L} E \left[\|\nabla F(\tilde{w}_{k-1})\|^2\right] + 2\gamma_k L\omega_{\text{dw}} E \left[\|w_{k-1} - H_{k-1}\|^2\right] + \gamma_k^2 \sigma^2(1 + \omega_{\text{up}}/N) \tag{S19}
$$

equivalent to:

$$
E \left[\|w_k - w_*\|^2\right] \leq (1 - \gamma_k \mu) \|w_{k-1} - w_*\|^2 - \gamma_k E \left[F(\tilde{w}_{k-1}) - F(w_*)\right] - \frac{\gamma_k}{2L} E \left[\|\nabla F(\tilde{w}_{k-1})\|^2\right] + 2\gamma_k L\omega_{\text{dw}} E \left[\|w_{k-1} - H_{k-1}\|^2\right] + \gamma_k^2 \sigma^2(1 + \omega_{\text{up}}) \tag{S20}
$$

Theorem 3 adapted to the case of decaying steps gives:

$$
E [Y_k] \leq \left(1 - \frac{\alpha_{\text{dw}}}{2}\right) E [Y_{k-1}] + 2\gamma_k^2 \left(\frac{1}{\alpha_{\text{dw}}} + \frac{\omega_{\text{up}}}{N}\right) E \left[\|\nabla F(\tilde{w}_{k-1})\|^2\right] + \frac{2\gamma_k^2 \sigma^2(1 + \omega_{\text{up}})}{Nb} \tag{S20}
$$

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

$$
E \left[\|w_k - w_*\|^2\right] + \gamma_k LCE [Y_k] \leq (1 - \gamma_k \mu) \|w_{k-1} - w_*\|^2 - \gamma_k E \left[F(\tilde{w}_{k-1}) - F(w_*)\right] - \frac{\gamma_k}{2L} E \left[\|\nabla F(\tilde{w}_{k-1})\|^2\right] + 2\gamma_k L\omega_{\text{dw}} E \left[\|w_{k-1} - H_{k-1}\|^2\right] + \gamma_k^2 \sigma^2(1 + \omega_{\text{up}}/Nb) + \left(1 - \frac{\alpha_{\text{dw}}}{2}\right) \gamma_k LCE [Y_{k-1}] + 2\gamma_k^2 LC \left(\frac{1}{\alpha_{\text{dw}}} + \frac{\omega_{\text{up}}}{N}\right) E \left[\|\nabla F(\tilde{w}_{k-1})\|^2\right] + \gamma_k^2 \sigma^2(1 + \omega_{\text{up}}/Nb)
$$

and reordering the terms gives:

$$
V_k \leq (1 - \gamma_k \mu) \|w_{k-1} - w_*\|^2 - \gamma_k E \left[F(\tilde{w}_{k-1}) - F(w_*)\right] + \left(1 - \frac{\alpha_{\text{dw}}}{2} + \frac{2\omega_{\text{dw}}}{C}\right) \gamma_k LCE \left[\|w_{k-1} - H_{k-1}\|^2\right] + \left(2\gamma_k^3 LC \left(\frac{1}{\alpha_{\text{dw}}} + \frac{\omega_{\text{up}}}{N}\right) - \frac{\gamma_k}{2L}\right) E \left[\|\nabla F(\tilde{w}_{k-1})\|^2\right] + (2\gamma_k LC + 1) \gamma_k^2 \sigma^2(1 + \omega_{\text{up}}/Nb),
$$

To reach a $(1 - \gamma \mu)$-convergence we first need $(1 - \frac{\alpha_{\text{dw}}}{2} + \frac{2\omega_{\text{dw}}}{C}) \gamma_k LC \leq (1 - \gamma_k \mu / \gamma_k) LC$

i.e $1 - \frac{\alpha_{\text{dw}}}{2} + \frac{2\omega_{\text{dw}}}{C} \leq (1 - \gamma_k \mu / \gamma_k) LC$.
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:

\[
\frac{\gamma_k}{\gamma_{k-1}} = \frac{\mu k - \mu + \bar{L}}{\mu k + L} = 1 - \frac{\mu}{\mu k + L}
\]

and so, the inequality is always true.

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

Secondly, it requires:

\[
2\gamma_k^2 C \left( \frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N} \right) = \frac{\gamma_k}{2L} \leq 0 \quad \iff \quad \gamma_k L \leq \frac{1}{4C \left( \frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N} \right)}
\]

\[
\iff \gamma_k L \leq \frac{1}{4\sqrt{\omega_{\text{dwn}} \left( \frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N} \right)}}
\]

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

\[
V_k \leq (1 - \gamma_k \mu) V_{k-1} - \gamma_k E [F(\tilde{w}_{k-1}) - F(w_\ast)] + \frac{\gamma_k^2 \sigma^2 \Phi(\gamma_k)}{Nb}
\]

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

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 \( (\lambda_k)_{k=1}^K \):

\[
\frac{1}{\sum_{k=1}^K \lambda_k} \sum_{k=1}^K \lambda_k E [F(\tilde{w}_{k-1}) - F(w_\ast)] \leq \frac{1}{\sum_{k=1}^K \lambda_k} \sum_{k=1}^K \frac{(1 - \gamma_k \mu) \lambda_k}{\gamma_k} V_{k-1} - \frac{\lambda_k V_k}{\gamma_k} + \frac{1}{\sum_{k=1}^K \lambda_k} \sum_{k=1}^K \lambda_k \frac{\gamma_k \sigma^2 \Phi(\gamma_k)}{Nb}
\]

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

\[
\frac{1}{\sum_{k=1}^K \lambda_k} \sum_{k=1}^K \lambda_k E [F(\tilde{w}_{k-1}) - F(w_\ast)] \leq \frac{1}{\sum_{k=1}^K \lambda_k} \sum_{k=1}^K \frac{1}{\gamma_k - 2\gamma_k - 1} V_{k-1} - \frac{1}{\gamma_k \gamma_{k-1}} V_k + \frac{1}{\sum_{k=1}^K \lambda_k} \sum_{k=1}^K \lambda_k \frac{\gamma_k \sigma^2 \Phi(\gamma_k)}{Nb}
\]

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

\[
\frac{1}{\sum_{k=1}^K \lambda_k} \sum_{k=1}^K \lambda_k E [F(\tilde{w}_{k-1}) - F(w_\ast)] \leq \frac{V_0}{\gamma_0 \gamma - 1 \mu K^2} + \frac{4 \mu K^2}{\sum_{k=1}^K \lambda_k \gamma_k \sigma^2 \Phi(\gamma_k)}.
\]
At the end, using the Jensen inequality - $E \left[ F(w_{k-1}) | w_{k-1} \right] \leq E [F(w_{k-1})]$, see Equation (S7) - we have for all $K$ in $\mathbb{N}$:

$$
\frac{1}{\sum_{k=1}^{K} \lambda_k} \sum_{k=1}^{K} \lambda_k E [F(w_{k-1}) - F(w_*)] \\
\leq \frac{V_0}{\gamma K \alpha_{\text{dwn}}} + 4 \frac{K}{\alpha_{\text{dwn}}} \sum_{k=1}^{K} \gamma_{k-1} \left( 1 + \frac{8 \gamma_k \omega_{\text{dwn}}}{\alpha_{\text{dwn}}} \right) \frac{\gamma_k \sigma^2 (1 + \omega_{\text{up}})}{N_b} \\
\leq \frac{V_0}{\gamma K \alpha_{\text{dwn}}} + 4 \frac{K}{\alpha_{\text{dwn}}} \sum_{k=1}^{K} \left( 1 + \frac{8 \gamma_k \omega_{\text{dwn}}}{\alpha_{\text{dwn}}} \right) \frac{\sigma^2 (1 + \omega_{\text{up}})}{N_b},
$$

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

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

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

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

and the demonstration is completed.

\[ \square \]

### E.4 Non-convex case (extra theorem)

In this section, we detail the convergence guarantee given for $\text{MC}$ 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 $\text{Ghost}$ and $\text{MC}$ 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 $\text{MC}$ in the non-convex case)

Under Assumptions 1, 2 and 4 (non-convex case), for a learning rate $\alpha_{\text{dwn}} = \frac{1}{8 \omega_{\text{dwn}}}$, for any step size $\gamma$ s.t.

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

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

$$
\frac{1}{K} \sum_{k=1}^{K} E \left[ \| \nabla F(w_{k-1}) \|^2 \right] \leq \frac{2 \left( F(w_0) - F(w_*) \right)}{\gamma K} + \frac{\gamma L \sigma^2 \Phi_{\text{non-convex}}}{N_b}.
$$

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

$$
\frac{1}{K} \sum_{k=1}^{K} E \left[ \| \nabla F(w_{k-1}) \|^2 \right] \leq 2 \sqrt{\frac{2 \sigma^2 (1 + \omega_{\text{up}}) \left( F(w_0) - F(w_*) \right)}{N_b K}} + O(K^{-1}).
$$
Proof. Let $k$ in $\mathbb{N}^*$, then smoothness (see Assumption 2) implies:

$$F(w_k) \leq F(w_{k-1}) + \langle \nabla F(w_{k-1}), w_k - w_{k-1} \rangle + \frac{L}{2}\|w_k - w_{k-1}\|^2$$

$$\iff F(w_k) \leq F(w_{k-1}) - \gamma \langle \nabla F(w_{k-1}), \tilde{g}_k \rangle + \frac{\gamma^2 L}{2}\|\tilde{g}_k\|^2.$$

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

$$-\mathbb{E} \langle \nabla F(w_{k-1}), \tilde{g}_k \rangle | w_{k-1} \rangle = -\langle \nabla F(w_{k-1}), \mathbb{E} \nabla F(\tilde{w}_{k-1}) | w_{k-1} \rangle \rangle
$$

$$= \frac{1}{2} \left( -\|\nabla F(w_{k-1})\|^2 - \mathbb{E} \left[ \|\nabla F(\tilde{w}_{k-1})\|^2 | w_{k-1} \right] \right) + \mathbb{E} \left[ \|\nabla F(w_{k-1}) - \nabla F(\tilde{w}_{k-1})\|^2 | w_{k-1} \right].$$

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

$$-\mathbb{E} \langle \nabla F(w_{k-1}), \tilde{g}_k \rangle | w_{k-1} \rangle \leq \frac{1}{2} \left( -\|\nabla F(w_{k-1})\|^2 - \mathbb{E} \left[ \|\nabla F(\tilde{w}_{k-1})\|^2 | w_{k-1} \right] \right) + L^2 \mathbb{E} \left[ \|w_{k-1} - \tilde{w}_{k-1}\|^2 | w_{k-1} \right].$$

Combining with Lemma S3, we obtain:

$$F(w_k) \leq F(w_{k-1}) - \frac{\gamma}{2}\|\nabla F(w_{k-1})\|^2 - \frac{\gamma}{2}\mathbb{E} \left[ \|\nabla F(\tilde{w}_{k-1})\|^2 | w_{k-1} \right] + \frac{\gamma L^2}{2} \mathbb{E} \left[ \|w_{k-1} - \tilde{w}_{k-1}\|^2 | w_{k-1} \right]
$$

$$+ \frac{\gamma^2 L}{2} \left( 1 + \frac{\omega_{up}}{N} \right) \|\nabla F(\tilde{w}_{k-1})\|^2 + \frac{\sigma^2(1 + \omega_{up})}{N^b} \right).$$

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

$$\mathbb{E} [F(w_k)] \leq \mathbb{E} [F(w_{k-1})] - \frac{\gamma}{2}\mathbb{E} \left[ \|\nabla F(w_{k-1})\|^2 \right] - \frac{\gamma}{2} \left( 1 - \frac{\gamma L}{1 + \omega_{up}/N} \right) \mathbb{E} \left[ \|\nabla F(\tilde{w}_{k-1})\|^2 \right]
$$

$$+ \frac{\gamma L^2}{2} \mathbb{E} \left[ \|w_{k-1} - \tilde{w}_{k-1}\|^2 \right] + \frac{\gamma^2 L}{2} \left( 1 + \frac{\omega_{up}}{N^b} \right) \|\nabla F(\tilde{w}_{k-1})\|^2.$$ 

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

$$\mathbb{E} [F(w_k)] \leq \mathbb{E} [F(w_{k-1})] - \frac{\gamma}{2} \mathbb{E} \left[ \|\nabla F(w_{k-1})\|^2 \right] - \frac{\gamma}{4} \mathbb{E} \left[ \|\nabla F(\tilde{w}_{k-1})\|^2 \right]
$$

$$+ \frac{\omega_{dwn} \gamma L^2}{2} \mathbb{E} \left[ \bar{\gamma}_k \right] + \frac{\gamma^2 L}{2} \left( 1 + \frac{\omega_{up}}{N^b} \right) \|\nabla F(\tilde{w}_{k-1})\|^2.$$ (S22)

Next, Theorem 3 gives:

$$\mathbb{E} [\bar{\gamma}_k] \leq \left( 1 - \frac{\alpha_{dwn}}{2} \right) \mathbb{E} [\bar{\gamma}_{k-1}] + 2\gamma^2 \left( \frac{1}{\alpha_{dwn}} + \frac{\omega_{up}}{N} \right) \mathbb{E} \left[ \|\nabla F(\tilde{w}_{k-1})\|^2 \right] + \frac{2\gamma^2 \sigma^2(1 + \omega_{up})}{N^b}.$$ 

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

$$\mathbb{E} [\bar{\gamma}_k] \leq \left( 1 - \frac{\alpha_{dwn}}{2} \right) \mathbb{E} [\bar{\gamma}_0] + 2\gamma^2 \left( \frac{1}{\alpha_{dwn}} + \frac{\omega_{up}}{N} \right) \sum_{t=1}^{k} \left( 1 - \frac{\alpha}{2} \right)^{k-t} \mathbb{E} \left[ \|\nabla F(\tilde{w}_{t-1})\|^2 \right]
$$

$$+ \frac{4\gamma^2 \sigma^2(1 + \omega_{up})}{\alpha_{dwn} N^b}.$$
where we considered for the last term of the above equation that \( \sum_{t=1}^k \left(1 - \frac{\alpha_{\text{dwn}}}{2}\right)^{k-t} \leq \frac{2}{\alpha_{\text{dwn}}} \). This is equivalent to:

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

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

\[
\frac{\gamma}{2} \mathbb{E} \left[\|\nabla F(w_{k-1})\|^2\right] \leq \mathbb{E} [F(w_{k-1}) - F(w_k)] - \frac{\gamma}{4} \mathbb{E} \left[\|\nabla F(\hat{w}_{k-1})\|^2\right] + \frac{\gamma L^2}{2} \left(\frac{4\omega_{\text{dwn}} \gamma \sigma^2 (1 + \omega_{\text{up}})}{N b \alpha_{\text{dwn}}^2}\right)
\]

\[
+ 2\omega_{\text{dwn}} \gamma \left(\frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N}\right) \sum_{t=1}^k \left(1 - \frac{\alpha_{\text{dwn}}}{2}\right)^{k-t} \mathbb{E} \left[\|\nabla F(\hat{w}_{t-1})\|^2\right]
\]

\[
+ \frac{\gamma^2 L}{2} \times \frac{\sigma^2 (1 + \omega_{\text{up}})}{N b}
\]

\[
\leq \mathbb{E} [F(w_{k-1}) - F(w_k)] - \frac{\gamma}{4} \mathbb{E} \left[\|\nabla F(\hat{w}_{k-1})\|^2\right] + \gamma^3 L^2 \omega_{\text{dwn}} \left(\frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N}\right) \sum_{k=1}^K \sum_{t=1}^k \left(1 - \frac{\alpha_{\text{dwn}}}{2}\right)^{k-t} \mathbb{E} \left[\|\nabla F(\hat{w}_{t-1})\|^2\right]
\]

\[
+ \frac{\gamma^2 \sigma^2 L (1 + \omega_{\text{up}})}{2 N b} \left(1 + \frac{4\gamma L \omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\right) K.
\]

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

\[
\frac{\gamma}{2} \sum_{k=1}^K \mathbb{E} \left[\|\nabla F(w_{k-1})\|^2\right] \leq \mathbb{E} [F(w_0) - F(w_k)] - \frac{\gamma}{4} \sum_{k=1}^K \mathbb{E} \left[\|\nabla F(\hat{w}_{k-1})\|^2\right]
\]

\[
+ \gamma^3 L^2 \omega_{\text{dwn}} \left(\frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N}\right) \sum_{k=1}^K \sum_{t=1}^k \left(1 - \frac{\alpha_{\text{dwn}}}{2}\right)^{k-t} \mathbb{E} \left[\|\nabla F(\hat{w}_{t-1})\|^2\right]
\]

\[
+ \frac{\gamma^2 \sigma^2 L (1 + \omega_{\text{up}})}{2 N b} \left(1 + \frac{4\gamma L \omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\right) K.
\]

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

\[
\frac{\gamma}{2} \sum_{k=1}^K \mathbb{E} \left[\|\nabla F(w_{k-1})\|^2\right] \leq \gamma F(w_0) - F(w_k) - \frac{\gamma}{4} \sum_{k=1}^K \mathbb{E} \left[\|\nabla F(\hat{w}_{k-1})\|^2\right]
\]

\[
+ \frac{2}{\alpha_{\text{dwn}}} \times \gamma^3 L^2 \omega_{\text{dwn}} \left(\frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N}\right) \sum_{k=1}^K \mathbb{E} \left[\|\nabla F(\hat{w}_{k-1})\|^2\right]
\]

\[
+ \frac{\gamma^2 \sigma^2 L (1 + \omega_{\text{up}})}{2 N b} \left(1 + \frac{4\gamma L \omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\right) K
\]

\[
\leq \mathbb{E} [F(w_0) - F(w_k)] + \left(\frac{2\gamma^3 L^2 \omega_{\text{dwn}}}{\alpha_{\text{dwn}}} \left(\frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N}\right) - \frac{\gamma}{4}\right) \sum_{k=1}^K \mathbb{E} \left[\|\nabla F(\hat{w}_{k-1})\|^2\right]
\]

\[
+ \frac{\gamma^2 \sigma^2 L (1 + \omega_{\text{up}})}{2 N b} \left(1 + \frac{4\gamma L \omega_{\text{dwn}}}{\alpha_{\text{dwn}}}\right) K.
\]
Now we consider that $2\gamma^3 F^2 (\frac{\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}) \left( \frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\up}}{N} \right) \leq \gamma/4$, and because for all $k \in \mathbb{N}$, $F(w_0) - F(w) \leq F(w_0) - F(w)$:

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

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

and $\alpha_{\text{dwn}} \leq \frac{1}{\delta \omega_{\text{dwn}}}$, we have:

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

denoting $\Phi_{\text{cvx}}(\gamma) := (1 + \omega_{\up}) \left( 1 + \frac{4\gamma L\omega_{\text{dwn}}}{\alpha_{\text{dwn}}} \right)$.

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

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

\[\square\]

### 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 $\llbracket 1, N \rrbracket$ holds a unique model $\hat{w}_{k-1}^i$.

For $k$ in $\mathbb{N}$:

1. $\bar{g}_k$ is now defined as $\bar{g}_k = \frac{1}{N} \sum_{i=1}^{N} \bar{g}_k^i (\hat{w}_{k-1}^i)$,
2. for all $i$ in $\llbracket 1, N \rrbracket$, $\bar{g}_k^i (\hat{w}_{k-1}^i)$ and $\nabla F(\hat{w}_{k-1}^i)$ must be replaced by $\bar{g}_k^i (\hat{w}_{k-1}^i)$ and $\nabla F(\hat{w}_{k-1}^i)$,
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 $\mathbb{N}^*$, $\bar{g}_k = \frac{1}{N} \sum_{i=1}^{N} \bar{g}_k^i (\hat{w}_{k-1}^i)$.

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 $\frac{1}{N} \sum_{i=1}^{N} \|w_{k-1} - H_{k-2}^i\|^2$, while we should consider the quantity $\frac{1}{N} \sum_{i=1}^{N} \|w_{k-1} - H_{k-2}\|^2$ for MCM. However this notation is correct considering that for MCM, for all $i$ in $\llbracket 1, N \rrbracket$, $H_k^i = H_k$. And it follows that we have $\frac{1}{N} \sum_{i=1}^{N} \|w_{k-1} - H_{k-2}^i\|^2 = \frac{1}{N} \sum_{i=1}^{N} \|w_{k-1} - H_{k-2}\|^2$. 

42
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 \( \frac{1}{N} \sum_{i=1}^{N} \nabla F(\hat{w}_{k-1}) - \nabla F(w_{k-1}) \) for \( k \in \mathbb{N}^* \). For this purpose we define a constant \( C \) such that \( C = 1 \) in the MCM-case and \( C = N \) in the Rand-MCM-case.

The proofs for quadratic functions rely on the fact that for any \( k \in \mathbb{N}^* \), \( \mathbb{E}[\nabla F(\hat{w}_{k-1} | w_{k-1}) = \nabla F(w_{k-1})] \).

**Definition 2** (Quadratic function). A function \( f : \mathbb{R}^d \to \mathbb{R} \) 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 tighter control of \( \left\| \sum_{i=1}^{N} \nabla F(\hat{w}_{k-1}^i) - \nabla F(w_{k-1}) \right\|^2 \).

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

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

**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 \( \hat{w}_{k-1}^i \) equal to \( \hat{w}_{k-1} \) for all \( i \) in \([1, N]\).

In the MCM-case:

\[
\mathbb{E}\left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla F(\hat{w}_{k-1}^i) - \nabla F(w_{k-1}) \right\|^2 \mid w_{k-1} \right] = \mathbb{E}[\nabla F(\hat{w}_{k-1}) - F(w_{k-1}) \mid w_{k-1}]
\]

\[
\leq L^2 \mathbb{E}\left[ \left\| \hat{w}_{k-1} - w_{k-1} \right\|^2 \mid w_{k-1} \right]
\]

\[
\leq L^2 \omega_{dwn} \left\| \Omega_{k-1} \right\|^2
\]

\[
\leq L^2 \omega_{dwn} \frac{1}{N} \sum_{i=1}^{N} \left\| w_{k-1} - H_{k-2} \right\|^2 ,
\]

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:

\[
\mathbb{E}\left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla F(\hat{w}_{k-1}^i) - \nabla F(w_{k-1}) \right\|^2 \mid w_{k-1} \right] = \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{E}\left[ \left\| \nabla F(\hat{w}_{k-1}^i) - \nabla F(w_{k-1}) \right\|^2 \mid w_{k-1} \right]
\]

\[
\leq \frac{L^2}{N^2} \sum_{i=1}^{N} \left\| \hat{w}_{k-1}^i - w_{k-1} \right\|^2
\]

\[
\leq \frac{L^2 \omega_{dwn}}{N} \times \frac{1}{N} \sum_{i=1}^{N} \left\| w_{k-1} - H_{k-2}^i \right\|^2
\]

\[
\leq \frac{L^2 \omega_{dwn}}{N} \frac{1}{N} \sum_{i=1}^{N} \left\| w_{k-1} - H_{k-2}^i \right\|^2 .
\]
We factorize the two results and define \( C \) such that \( C = 1 \) in the MCM-case and \( C = N \) in the Rand-MCM-case, and the result follows.

\[
\mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla F(\hat{w}_k^{i} - 1) - \nabla F(w_{k-1}) \right\|^2 \bigg| w_{k-1} \right] \leq \frac{L^2 \omega\text{down}}{C} \frac{1}{N} \sum_{i=1}^{N} \left\| w_{k-1} - H^i_{k-2} \right\|^2 .
\]

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. \( \hat{w}_{k-1} \) while here we take the expectation w.r.t. \( w_{k-1} \). This is because we remove \( \hat{w}_{k-1} \) from the gradient and give a result which depends of \( \left\| \nabla F(w_{k-1}) \right\|^2 \) instead of \( \left\| \nabla F(\hat{w}_{k-1}) \right\|^2 \). This is made possible by the fact that for all \( k \) in \( \mathbb{N} \), for quadratic functions, we have

\[
\mathbb{E} \left[ \nabla F(\hat{w}_k^{i} - 1) \right] = \nabla F(w_{k-1}).
\]

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

\[
\mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \hat{g}^i_k(\hat{w}_k^{i} - 1) - \nabla F(\hat{w}_k^{i} - 1) \right\|^2 \bigg| w_{k-1} \right] \leq \frac{\omega\text{up}}{N} \left\| \nabla F(w_{k-1}) \right\|^2 + \frac{\sigma^2 (1 + \omega\text{up})}{Nb} \\
+ \frac{\omega\text{up} \omega\text{down} L^2}{N} \frac{1}{N} \sum_{i=1}^{N} \left\| w_{k-1} - H^i_{k-2} \right\|^2 ,
\]

(S23)

\[
\mathbb{E} \left[ \left\| \tilde{g}_k \right\|^2 \bigg| w_{k-1} \right] \leq \left( 1 + \frac{\omega\text{up}}{N} \right) \left\| \nabla F(w_{k-1}) \right\|^2 + \frac{\sigma^2 (1 + \omega\text{up})}{Nb} \\
+ L^2 \omega\text{down} \left( \frac{1}{C} + \frac{\omega\text{up}}{N} \right) \frac{1}{N} \sum_{i=1}^{N} \left\| w_{k-1} - H^i_{k-2} \right\|^2 .
\]

(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:

\[
\mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \hat{g}^i_k(\hat{w}_k^{i} - 1) - \nabla F(\hat{w}_k^{i} - 1) \right\|^2 \bigg| w_{k-1} \right] \\
= \mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \left( \hat{g}^i_k(\hat{w}_k^{i} - 1) - g^i_k(\hat{w}_k^{i} - 1) \right) \right\|^2 \bigg| w_{k-1} \right] \\
+ \mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \left( g^i_k(\hat{w}_k^{i} - 1) - \nabla F(\hat{w}_k^{i} - 1) \right) \right\|^2 \bigg| w_{k-1} \right] ,
\]
the first term is bounded by Assumption 1 and the last term by Assumption 4:

\[
\mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \left( \hat{g}_k^i (\hat{\omega}_{k-1}^i) - \nabla F(\hat{\omega}_{k-1}^i) \right) \right\|^2 \bigg| w_{k-1} \right] \\
\leq \frac{\omega_{up}}{N^2} \sum_{i=1}^{N} \mathbb{E} \left[ \left\| \hat{g}_k^i (\hat{\omega}_{k-1}^i) \right\|^2 \bigg| w_{k-1} \right] + \frac{\sigma^2}{N} \\
\leq \frac{\omega_{up}}{N^2} \sum_{i=1}^{N} \mathbb{E} \left[ \left\| \hat{g}_k^i (\hat{\omega}_{k-1}^i) - \nabla F(\hat{\omega}_{k-1}^i) \right\|^2 \bigg| w_{k-1} \right] \\
+ \frac{\omega_{up}}{N^2} \sum_{i=1}^{N} \mathbb{E} \left[ \left\| \nabla F(\hat{\omega}_{k-1}^i) \right\|^2 \bigg| w_{k-1} \right] + \frac{\sigma^2}{N}.
\]

And again applying Assumption 4 on \( \mathbb{E} \left[ \left\| \hat{g}_k^i (\hat{\omega}_{k-1}^i) - \nabla F(\hat{\omega}_{k-1}^i) \right\|^2 \bigg| w_{k-1} \right] \) for \( i \) in \( \{1, \ldots, N\} \):

\[
\mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \left( \hat{g}_k^i (\hat{\omega}_{k-1}^i) - \nabla F(\hat{\omega}_{k-1}^i) \right) \right\|^2 \bigg| w_{k-1} \right] = \frac{\omega_{up}}{N^2} \sum_{i=1}^{N} \mathbb{E} \left[ \left\| \nabla F(\hat{\omega}_{k-1}^i) \right\|^2 \bigg| w_{k-1} \right] \\
+ \frac{\omega_{up}}{N^2} \sum_{i=1}^{N} \mathbb{E} \left[ \left\| \nabla F(\hat{w}_{k-1}) \right\|^2 \bigg| w_{k-1} \right],
\]

Now, we have:

\[
\frac{\omega_{up}}{N^2} \sum_{i=1}^{N} \mathbb{E} \left[ \left\| \nabla F(\hat{\omega}_{k-1}^i) \right\|^2 \bigg| w_{k-1} \right] = \frac{\omega_{up}}{N^2} \sum_{i=1}^{N} \mathbb{E} \left[ \left\| \nabla F(\hat{\omega}_{k-1}^i) - \nabla F(w_{k-1}) \right\|^2 \bigg| w_{k-1} \right] \\
+ \frac{\omega_{up}}{N^2} \sum_{i=1}^{N} \mathbb{E} \left[ \left\| \nabla F(w_{k-1}) \right\|^2 \bigg| w_{k-1} \right],
\]

using smoothness (Assumption 2) gives:

\[
\frac{\omega_{up}}{N^2} \sum_{i=1}^{N} \mathbb{E} \left[ \left\| \nabla F(\hat{\omega}_{k-1}^i) \right\|^2 \bigg| w_{k-1} \right] = \frac{\omega_{up} \omega_{dwn} L^2}{N} \frac{1}{N} \sum_{i=1}^{N} \left\| w_{k-1} - H_{k-2}^i \right\|^2 + \frac{\omega_{up}}{N} \left\| \nabla F(w_{k-1}) \right\|^2,
\]

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

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

\[
\mathbb{E} \left[ \left( \frac{1}{N} \sum_{i=1}^{N} \hat{g}_k^i (\hat{\omega}_{k-1}^i) \right)^2 \bigg| w_{k-1} \right] = \mathbb{E} \left[ \left( \frac{1}{N} \sum_{i=1}^{N} \hat{g}_k^i (\hat{\omega}_{k-1}^i) - \nabla F(w_{k-1}) \right)^2 \bigg| w_{k-1} \right] \\
+ \left\| \nabla F(w_{k-1}) \right\|^2 \\
= \mathbb{E} \left[ \left( \frac{1}{N} \sum_{i=1}^{N} \hat{g}_k^i (\hat{\omega}_{k-1}^i) - \nabla F(\hat{\omega}_{k-1}^i) \right)^2 \bigg| w_{k-1} \right] \\
+ \mathbb{E} \left[ \left( \frac{1}{N} \sum_{i=1}^{N} \nabla F(\hat{\omega}_{k-1}^i) - \nabla F(w_{k-1}) \right)^2 \bigg| w_{k-1} \right] \\
+ \left\| \nabla F(w_{k-1}) \right\|^2.
\]

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

\[
\mathbb{E} \left[ \left( \frac{1}{N} \sum_{i=1}^{N} \nabla F(\hat{\omega}_{k-1}^i) - \nabla F(w_{k-1}) \right)^2 \bigg| w_{k-1} \right] \leq \frac{L^2 \omega_{dwn}}{C} \frac{1}{N} \sum_{i=1}^{N} \left\| w_{k-1} - H_{k-2}^i \right\|^2.
\]
Thus, we have:

\[
E \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \mathcal{g}_i^i (\tilde{w}_{k-1}^i) \right\|^2 \mid w_{k-1} \right] = E \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \mathcal{g}_i^i (\tilde{w}_{k-1}^i) - \nabla F (\tilde{w}_{k-1}^i) \right\|^2 \mid w_{k-1} \right]
\]

\[
+ \frac{\omega_{\text{dwn}} L^2}{C} \frac{1}{N} \sum_{i=1}^{N} \left\| w_{k-1} - H_{k-2}^i \right\|^2 + \left\| \nabla F (w_{k-1}) \right\|^2,
\]

and eq. (S23) allows to conclude.

\[\square\]

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 \((w_i)_{i \in \mathbb{N}}\). 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 \(||\nabla F(\tilde{w}_{k-1})||^2\) 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 \(E[\nabla F(\tilde{w}_{k-1})] = \nabla F(w_{k-1})\). 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 \(E[\tilde{g}_k]\) 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 \(\gamma \leq \frac{1}{8L\omega_{\text{dwn}} \sqrt{(1/C + \omega_{\text{up}}/N)}}\) and \(\alpha_{\text{dwn}} \leq 1/(8\omega_{\text{dwn}})\), then for all \(k \in \mathbb{N}\):

\[
\begin{align*}
\frac{1}{N} \sum_{i=1}^{N} E \left[ \left\| w_k - H_{k-1}^i \right\|^2 \mid w_{k-1} \right] & \leq 2\gamma^2 \left( \frac{1}{\alpha_{\text{dwn}}} + \frac{\omega_{\text{up}}}{N} \right) \sum_{t=1}^{k} (1 - \frac{\alpha_{\text{dwn}}}{2})^{k-t} E \left[ \left\| \nabla F(w_{t-1}) \right\|^2 \mid w_{t-1} \right] \\
& + \frac{4\gamma^2 \sigma^2 (1 + \omega_{\text{up}})}{\alpha_{\text{dwn}} N b}.
\end{align*}
\]

**Proof.** Let \(k \in \mathbb{N}^+\) and \(i \in \{1, \ldots, N\}\), from Theorem S8 we have:

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

with

\[
\begin{align*}
\text{Var}_1 &= E \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \tilde{g}_i^i (\tilde{w}_{k-1}^i) + E \left[ \nabla F(\tilde{w}_{k-1}^i) \mid w_{k-1} \right] \right\|^2 \mid w_{k-1} \right] \\
\text{Var}_2 &= \omega_{\text{dwn}} \frac{1}{N} \sum_{i=1}^{N} \left\| w_{k-1} - H_{k-2}^i \right\|^2 \\
\text{Bias}^2 &= \left\| E \left[ w_k - H_{k-1}^i \mid w_{k-1} \right] \right\|^2.
\end{align*}
\]

46
Recall that in the case of quadratic functions, we have for all $i \in [1, N]$: $E \left[ \nabla F(\hat{w}_{k-1}^i) \mid w_{k-1} \right] = \nabla F(w_{k-1})$. And so for the first term of variance we can decompose as following:

$$
\text{Var}_1 = E \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_k(\hat{w}_{k-1}^i) - \nabla F(\hat{w}_{k-1}^i) \right\| \mid w_{k-1} \right]^2 w_{k-1}
= E \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_k(\hat{w}_{k-1}^i) - \nabla F(w_{k-1}) \right\| \mid w_{k-1} \right]^2 w_{k-1}
= E \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_k(\hat{w}_{k-1}^i) - \nabla F(w_{k-1}) \right\| \mid w_{k-1} \right]^2 w_{k-1}
+ E \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla F(\hat{w}_{k-1}^i) - \nabla F(w_{k-1}) \right\| \mid w_{k-1} \right]^2 w_{k-1}.
$$

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

$$
E \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_k(\hat{w}_{k-1}^i) - \nabla F(\hat{w}_{k-1}^i) \right\| \mid w_{k-1} \right]^2 w_{k-1} = \frac{\omega_{\text{up}} \omega_{\text{dwn}} L^2}{N} \frac{1}{N} \sum_{i=1}^{N} \|w_{k-1} - H_{k-2}^i\|^2
+ \frac{\lambda_{\text{up}}}{N} \|\nabla F(w_{k-1})\|^2
+ \frac{\sigma^2 (1 + \omega_{\text{up}})}{Nb},
$$

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

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

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

Now we focus on the squared bias $\text{Bias}^2$ exactly like in Theorem S8 and we obtain:

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

At the end:

$$
E \left[ \|w_k - H_{k-1}^i\|^2 \mid w_{k-1} \right] \leq 2\gamma^2 L^2 \omega_{\text{dwn}} \left( \frac{1}{C} + \frac{\omega_{\text{up}}}{N} \right) \frac{1}{N} \sum_{i=1}^{N} \|w_{k-1} - H_{k-2}^i\|^2
+ \gamma^2 (1 + \frac{1}{\alpha_{\text{dwn}}}) \frac{2\omega_{\text{up}}}{N} \|\nabla F(w_{k-1})\|^2
+ \frac{2\gamma^2 \omega_{\text{dwn}}}{N} \|w_{k-1} - H_{k-2}^i\|^2
+ \frac{2\gamma^2 \sigma^2 (1 + \omega_{\text{up}})}{Nb}.
$$
Summing this last equation over the $N$ devices gives:

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

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

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

Thus, we can write:

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

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

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

and the result follows.

\[ \square \]

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 $\alpha_{\text{dwn}} = 1/(8\omega_{\text{dwn}})$, after running $K$ iterations:

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

with

$$
\Phi = (1 + \omega_{\text{up}}) \left( 1 + 32\frac{\omega_{\text{dwn}}}{\sqrt{K}} \times \frac{1}{\sqrt{K}} \left( \frac{1}{C} + \frac{\omega_{\text{up}}}{N} \right) \right)
$$

and $C = N$ 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:

$$
\|w_k - w_*\|^2 \leq \|w_{k-1} - w_*\|^2 - 2\gamma \langle \bar{g}_k, w_{k-1} - w_* \rangle + \gamma^2 \|\bar{g}_k\|^2 .
$$
Because $F$ is quadratic, we have $E \left[ \nabla F(\tilde{w}_{k-1}) \mid w_{k-1} \right] = \nabla F(w_{k-1})$, thus taking expectation gives:

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

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

$$E \left[ \|w_k - w_*\|^2 \mid w_{k-1} \right] \leq \|w_{k-1} - w_*\|^2 - \gamma \left(F(w_{k-1}) - F(w_*) + \frac{1}{L} \|\nabla F(w_{k-1})\|^2\right) + \gamma^2 E \left[ \|\tilde{g}_k\|^2 \mid w_{k-1} \right].$$

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

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

which gives:

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

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

$$E \left[ \|w_k - w_*\|^2 \right] \leq E \left[ \|w_{k-1} - w_*\|^2 \right] - \gamma E \left[ F(w_{k-1}) - F(w_*) \right] - \frac{\gamma}{L} \left(1 - \frac{\omega_{up}}{N}\right) E \left[ \|\nabla F(w_{k-1})\|^2 \right] + \gamma^2 L^2 \omega_{dwn} \left(\frac{1}{C} + \frac{\omega_{up}}{N}\right) \times 2\gamma^2 \left(\frac{1}{\alpha_{dwn}} + \frac{\omega_{up}}{N}\right) \sum_{t=1}^{k} (1 - \frac{\alpha_{dwn}}{2})^{k-t} E \left[ \|\nabla F(w_{t-1})\|^2 \right] + \frac{\sigma^2\gamma^2(1 + \omega_{up})}{Nb} + \gamma^2 L^2 \omega_{dwn} \left(\frac{1}{C} + \frac{\omega_{up}}{N}\right) \times \frac{4\sigma^2\gamma^2(1 + \omega_{up})}{\alpha_{dwn} Nb}.$$
Next, we consider - as in previous proofs - that $\gamma L (1 + \omega_{up}/N) \leq 1/2$, and thus $\frac{\gamma L}{2} \left(1 - \gamma L \left(1 + \frac{\omega_{up}}{N}\right)\right) \geq \frac{\gamma}{2}$. Next we carry out the “top-down recurrence”:

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

We invert the double-sum, it leads to:

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

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

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

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

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

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

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

denoting $\Phi^{\text{Rd}}(\gamma) = (1 + \omega_{up}) \left(1 + \frac{4\gamma^2 L^2 \omega_{dwn}}{K} \left(\frac{1}{C} + \frac{\omega_{up}}{N}\right)\right).$
G Adaptaotion to the heterogeneous scenario

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

We make the following assumption on the heterogeneity.

\[ \Phi \]

We denote (Bounded gradient at\footnote{NB: I need to understand this notation.}) Assumption 6 (Recursive inequalities over memory term) Lemma S8 term used in the uplink broadcasting can be bounded using a recursion:

\[ \text{gradients as defined in Assumptions 4 and 6 can be controlled as following:} \]

each node to the central server can be bounded as following:

\[ B \]

MCM and Lemma S8 corresponds to Lemma S13. Secondly, Appendix G.2 gives the demonstration of Heterog.

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 \( \hat{w}_{k-1} \); and Lemma S8 corresponds to Lemma S13. Secondly, Appendix G.2 gives the demonstration of Heterog.

We denote \( \Phi_{\text{Heterog}}(\gamma) := (1 + 8\omega_{up}) \left(1 + \frac{8\gamma L\omega_{dwn}}{\alpha_{dwn}}\right) \) and \( \gamma_{\text{Heterog}}^{\max} \) such that:

\[ \gamma_{\text{Heterog}}^{\max} L \leq \min \left\{ \gamma_{\text{max}}, \frac{1}{10}, \frac{1}{8\sqrt{2\omega_{dwn}/\omega_{up}} \cdot \omega_{up}/N} \right\}. \]

We make the following assumption on the heterogeneity.

**Assumption 6 (Bounded gradient at \( w_0 \)).** There is a constant \( B \) in \( \mathbb{R}_+ \), s.t.:

\[ \frac{1}{N} \sum_{i=0}^{N} \| \nabla F_i(w_i) \|^2 = B^2 \].

And we denote for all \( i \in [1, N] \), \( h_i^* = \nabla F_i(w_i) \).

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:

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

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

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

**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:

\[ \mathbb{E} \left[ \Xi_k \bigg| w_{k-1} \right] \leq (1 - \alpha_{up}) \mathbb{E} \left[ \Xi_{k-1} \bigg| w_{k-1} \right] + \frac{2\omega_{up}L}{N} \mathbb{E} \left[ \left\| \nabla F(\hat{w}_{k-1}), \hat{w}_{k-1} - w_* \right\| w_{k-1} \right] + \frac{2\sigma^2}{Nb} \alpha_{up}. \]

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

\[ \mathbb{E} \left[ \left\| \tilde{g}_k \right\|^2 \bigg| \hat{w}_{k-1} \right] \leq \frac{4\omega_{up}}{N} L \mathbb{E} \left[ \left\| \nabla F(\hat{w}_{k-1}), \hat{w}_{k-1} - w_* \right\| \hat{w}_{k-1} \right] + 2\omega_{up} \mathbb{E} \left[ \Xi_{k-1} \bigg| \hat{w}_{k-1} \right] + \frac{\sigma^2}{Nb} (1 + 4\omega_{up}), \]

\[ \mathbb{E} \left[ \left\| \tilde{g}_k - \nabla F(\hat{w}_{k-1}) \right\|^2 \bigg| \hat{w}_{k-1} \right] \leq \frac{4\omega_{up}L}{N} \mathbb{E} \left[ \left\| \nabla F(\hat{w}_{k-1}), \hat{w}_{k-1} - w_* \right\| \hat{w}_{k-1} \right] + 2\omega_{up} \mathbb{E} \left[ \Xi_{k-1} \bigg| \hat{w}_{k-1} \right] + \frac{\sigma^2}{Nb} (1 + 4\omega_{up}), \]

\[ 51 \]
Lemma S9 extends Lemma S3.

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

\[
\mathbb{E} \left[ \| \hat{g}_k \|^2 \mid \hat{w}_{k-1} \right] = \| \nabla F(\hat{w}_{k-1}) \|^2 + \mathbb{E} \left[ \| \hat{g}_k - \nabla F(\hat{w}_{k-1}) \| \mid \hat{w}_{k-1} \right]
\]

\[
\leq \mathbb{E} \left[ \langle \nabla F(\hat{w}_{k-1}), \hat{w}_{k-1} - w_s \rangle \mid \hat{w}_{k-1} \right] + \mathbb{E} \left[ \| \hat{g}_k - \nabla F(\hat{w}_{k-1}) \| \mid \hat{w}_{k-1} \right]
\]

Secondly:

\[
\mathbb{E} \left[ \| \hat{g}_k - \nabla F(\hat{w}_{k-1}) \|^2 \mid \hat{w}_{k-1} \right]
\]

\[
= \mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \left( \hat{\Delta}_{k-1}^i + h_{k-1}^i - \nabla F_i(\hat{w}_{k-1}) \right) \right\|^2 \mid \hat{w}_{k-1} \right]
\]

\[
= \mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \left( \hat{\Delta}_{k-1}^i + h_{k-1}^i - \partial F(\hat{w}_{k-1}) + g_k^i(\hat{w}_{k-1}) - \nabla F_i(\hat{w}_{k-1}) \right) \right\|^2 \mid \hat{w}_{k-1} \right]
\]

\[
= \mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \hat{\Delta}_{k-1}^i - \Delta_{k-1}^i \right\|^2 \mid \hat{w}_{k-1} \right] + \mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} g_k^i(\hat{w}_{k-1}) - \nabla F_i(\hat{w}_{k-1}) \right\|^2 \mid \hat{w}_{k-1} \right]
\]

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:

\[
\mathbb{E} \left[ \| \hat{g}_k - \nabla F(\hat{w}_{k-1}) \|^2 \mid \hat{w}_{k-1} \right] = \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{E} \left[ \left\| \hat{\Delta}_{k-1}^i - \Delta_{k-1}^i \right\|^2 \mid \hat{w}_{k-1} \right]
\]

\[
+ \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{E} \left[ \| g_k^i(\hat{w}_{k-1}) - \nabla F_i(\hat{w}_{k-1}) \|^2 \mid \hat{w}_{k-1} \right].
\]

Then, for any $i$ in $[1,N]$ as $\mathbb{E} \left[ \| \hat{\Delta}_{k-1}^i - \Delta_{k-1}^i \|^2 \mid \hat{w}_{k-1} \right] = 0$, and using Assumption 1 we have:

\[
\mathbb{E} \left[ \left\| \hat{\Delta}_{k-1}^i - \Delta_{k-1}^i \right\|^2 \mid \hat{w}_{k-1} \right] = 0
\]

\[
\mathbb{E} \left[ \| g_k^i(\hat{w}_{k-1}) - \nabla F_i(\hat{w}_{k-1}) \|^2 \mid \hat{w}_{k-1} \right]
\]

\[
\leq \frac{\omega_{up}}{N^2} \sum_{i=1}^{N} \mathbb{E} \left[ \left\| \Delta_{k-1}^i \right\|^2 \mid \hat{w}_{k-1} \right]
\]

\[
+ \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{E} \left[ \| g_k^i(\hat{w}_{k-1}) - \nabla F_i(\hat{w}_{k-1}) \|^2 \mid \hat{w}_{k-1} \right].
\]

Furthermore with Lemma S6 and Assumption 4:

\[
\mathbb{E} \left[ \| \hat{g}_k - \nabla F(\hat{w}_{k-1}) \|^2 \mid \hat{w}_{k-1} \right] \leq \frac{2\omega_{up}}{N^2} \sum_{i=1}^{N} \mathbb{E} \left[ \| g_k(\hat{w}_{k-1}) - h_i^i \|^2 \mid \hat{w}_{k-1} \right]
\]

\[
+ 2\omega_{up} \mathbb{E} \left[ \Xi_{k-1} \mid \hat{w}_{k-1} \right] + \frac{\sigma^2}{Nb}.
\]

And finally with Lemma S7:

\[
\mathbb{E} \left[ \| \hat{g}_k - \nabla F(\hat{w}_{k-1}) \|^2 \mid \hat{w}_{k-1} \right] \leq \frac{4\omega_{up} L}{N} \mathbb{E} \left[ \langle \nabla F(\hat{w}_{k-1}), \hat{w}_{k-1} - w_s \rangle \mid \hat{w}_{k-1} \right] + \frac{4\omega_{up} \sigma^2}{Nb}
\]

\[
+ 2\omega_{up} \mathbb{E} \left[ \Xi_{k-1} \mid \hat{w}_{k-1} \right] + \frac{\sigma^2}{Nb},
\]

from which we derive the two inequalities of the lemma. 

$\square$
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 \( \gamma \leq 1/(8\omega_{dwn}L) \) and \( \alpha_{dwn} \leq 1/(8\omega_{dwn}) \), then for all \( k \in \mathbb{N} \):

\[
\mathbb{E} [Y_k | w_{k-1}] \leq \left(1 - \frac{\alpha_{dwn}}{2}\right) Y_{k-1} + 2\gamma^2 L \left( \frac{1}{\alpha_{dwn}} + \frac{4\omega_{up}}{\mathbb{N}} \right) \mathbb{E} [\langle \nabla F(\hat{w}_{k-1}), \hat{w}_{k-1} - w^* \rangle | \hat{w}_{k-1}] + 4\gamma^2 \omega_{up} \mathbb{E} [\Xi_{k-1} | \hat{w}_{k-1}] + \frac{2\gamma^2 \sigma^2 (1 + 4\omega_{up})}{\mathbb{N}b}.
\]

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

\[
\begin{align*}
\Omega_k &= w_k - H_{k-1} \\
\hat{\Omega}_k &= C_{dwn}(\Omega_k) \\
\hat{w}_k &= \hat{\Omega}_k + H_{k-1}.
\end{align*}
\]

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

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

We first have:

\[
\text{Var}_{11} = \mathbb{E} \left[ \|\nabla F(\hat{w}_{k-1})\|^2 | w_{k-1} \right] = \mathbb{E} \left[ \mathbb{E} \left[ \|\nabla F(\hat{w}_{k-1})\|^2 | \hat{w}_{k-1} \right] | w_{k-1} \right],
\]

so, we can use Lemma S9:

\[
\text{Var}_{11} = \frac{4\omega_{up}L}{\mathbb{N}} \mathbb{E} [\langle \nabla F(\hat{w}_{k-1}), \hat{w}_{k-1} - w^* \rangle | \hat{w}_{k-1}] + 2\omega_{up} \mathbb{E} [\Xi_{k-1} | \hat{w}_{k-1}] + \frac{\sigma^2}{\mathbb{N}b}(1 + 4\omega_{up}).
\]

The other terms are exactly as before in Theorem S8:

\[
\begin{align*}
\text{Var}_{12} &\leq 2\omega_{dwn} Y_{k-1} \\
\text{Var}_2 &\leq \omega_{dwn} Y_{k-1} \\
\text{Bias}^2 &\leq (1 - \alpha_{dwn}) Y_{k-1} + \gamma^2 L(1 + \frac{1}{\alpha_{dwn}}) \mathbb{E} [\langle \nabla F(\hat{w}_{k-1}), \hat{w}_{k-1} - w^* \rangle | \hat{w}_{k-1}].
\end{align*}
\]

At the end:

\[
\begin{align*}
\mathbb{E} [Y_k | w_{k-1}] \leq & (1 - \alpha_{dwn}) Y_{k-1} + \gamma^2 L(1 + \frac{1}{\alpha_{dwn}}) \mathbb{E} [\langle \nabla F(\hat{w}_{k-1}), \hat{w}_{k-1} - w^* \rangle | \hat{w}_{k-1}] \\
&+ \frac{8\omega_{up} \gamma^2 L}{\mathbb{N}} \mathbb{E} [\langle \nabla F(\hat{w}_{k-1}), \hat{w}_{k-1} - w^* \rangle | \hat{w}_{k-1}] \\
&+ 4\gamma^2 \omega_{up} \mathbb{E} [\Xi_{k-1} | \hat{w}_{k-1}] + \frac{2\gamma^2 \sigma^2}{\mathbb{N}b}(1 + 4\omega_{up}) \\
&+ 2\gamma^2 L^2 \omega_{dwn} Y_{k-1} + 2\alpha_{dwn}^2 \omega_{dwn} Y_{k-1},
\end{align*}
\]

53
which is equivalent to:

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

Next, we require as in Theorem S8:

\[
\begin{align*}
2\alpha_{dwn}^2 \omega_{dwn} & \leq \frac{1}{4} \alpha_{dwn} \iff \alpha_{dwn} \leq \frac{1}{8\omega_{dwn}}, \\
2\gamma^2 L^2 \omega_{dwn} & \leq \frac{1}{4} \alpha_{dwn} = \frac{1}{8\omega_{dwn}}, \text{ by taking } \alpha_{dwn} = \frac{1}{8\omega_{dwn}} \iff \gamma \leq \frac{1}{8\omega_{dwn} L}, \\
1 + \frac{1}{\alpha_{dwn}} & \leq \frac{2}{\alpha_{dwn}} \text{ which is not restrictive if } \omega_{dwn} \geq 1,
\end{align*}
\]

and it leads to the final result taking unconditional expectation. \qed

### 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 \( \alpha_{dwn} \leq \frac{1}{8\omega_{dwn}} \) and \( \alpha_{up} (1 + \omega_{up}) \leq 1 \), taking a step size s.t. \( \gamma \leq \gamma_{\text{maxHeterog}} \), for any \( k \in \mathbb{N} \), defining:

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

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

\[
V_k \leq V_{k-1} - \gamma \mathbb{E}[F(\tilde{w}_{k-1}) - F(w_*)] + \frac{\gamma^2 \sigma^2 \mathcal{H}_{\text{Heterog}}(\gamma)}{Nb}.
\]

**Proof.** We denote for \( k \in \mathbb{N}^* \)

\[
\bar{g}_k = \frac{1}{N} \sum_{i=1}^N \bar{\Delta}_{k-1} + h_{k-1}^i \text{ with } \Delta_{k-1} = g^i_k(\bar{w}_{k-1}) - h_{k-1}^i,
\]

and

\[
\Xi_k = \frac{1}{N^2} \sum_{i=1}^N \mathbb{E}\left[\left\| h_{k-1}^i - \nabla F_i(w_*) \right\|^2 \mid \tilde{w}_{k-1}\right].
\]

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

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

Next, we expand the inner product as following:

\[
\left\| w_k - w_* \right\|^2 \leq \left\| w_{k-1} - w_* \right\|^2 - 2\gamma \langle \bar{g}_k, \tilde{w}_{k-1} - w_* \rangle - 2\gamma \langle \bar{g}_k, w_{k-1} - \tilde{w}_{k-1} \rangle + \gamma^2 \left\| \bar{g}_k \right\|^2.
\]

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

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

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

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

\[\text{54}\]
Now using Cauchy-Schwarz inequality (eq. (S5)) and smoothness:

\[- E \left[ 2 \gamma \langle \nabla F(\hat{w}_{k-1}) - \nabla F(w_{k-1}), w_{k-1} - \hat{w}_{k-1} \rangle \mid w_{k-1} \right] \]
\[= 2\gamma E \left[ \langle \nabla F(\hat{w}_{k-1}) - \nabla F(w_{k-1}), \hat{w}_{k-1} - w_{k-1} \rangle \mid w_{k-1} \right] \]
\[\leq 2\gamma L E \left[ \|\hat{w}_{k-1} - w_{k-1}\|^2 \mid w_{k-1} \right],\]

and thus:

\[E \left[ \|w_k - w_*\|^2 \mid w_{k-1} \right] \leq \|w_{k-1} - w_*\|^2 - 2\gamma E \left[ \langle \nabla F(\hat{w}_{k-1}), \hat{w}_{k-1} - w_* \rangle \mid w_{k-1} \right] \]
\[+ 2\gamma L E \left[ \|\hat{w}_{k-1} - w_{k-1}\|^2 \mid w_{k-1} \right] \]
\[+ \left( 1 + \frac{4\omega_{up}}{N} \right) \gamma^2 L E \left[ \langle \nabla F(\hat{w}_{k-1}), \hat{w}_{k-1} - w_* \rangle \mid \hat{w}_{k-1} \right] \]
\[+ 2\omega_{up} \gamma^2 E \left[ \|\hat{w}_{k-1}\| \mid \hat{w}_{k-1} \right] + \frac{\sigma^2 \gamma^2}{Nb} \left( 1 + 4\omega_{up} \right).\]

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

\[E \left[ \|w_k - w_*\|^2 \mid w_{k-1} \right] \leq \|w_{k-1} - w_*\|^2 - \gamma E \left[ \langle \nabla F(\hat{w}_{k-1}), \hat{w}_{k-1} - w_* \rangle \mid w_{k-1} \right] \]
\[+ 2\gamma L E \left[ \|\hat{w}_{k-1} - w_{k-1}\|^2 \mid w_{k-1} \right] \]
\[+ 2\omega_{up} \gamma^2 E \left[ \|\hat{w}_{k-1}\| \mid \hat{w}_{k-1} \right] + \frac{\sigma^2 \gamma^2}{Nb} \left( 1 + 4\omega_{up} \right).\]  

With Lemma S8, we have:

\[E \left[ \|w_k - w_*\| \mid w_{k-1} \right] \leq \left( 1 - \alpha_{up} \right) E \left[ \|w_{k-1} \mid w_{k-1} \right] \]
\[+ \frac{2\alpha_{up} L}{N} E \left[ \langle \nabla F(\hat{w}_{k-1}), \hat{w}_{k-1} - w_* \rangle \mid w_{k-1} \right] \]
\[+ \frac{2N}{\sigma^2} \sigma_{up}. \]  

and Theorem S14 gives:

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

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

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

with \(C_1 = 2\omega_{up} (1 + 8\gamma L \omega_{dwn}/\alpha_{dwn})/\alpha_{up}\) and \(C_2 = 4\omega_{dwn}/\alpha_{dwn}.\)
We combine previous equations as follows (S26) + $\gamma^2C_1(S27) + C_2(S28)$:

\[
\mathbb{E} \left[ \|w_k - w_*\|^2 \right] + \gamma^2C_1\mathbb{E} [\xi_k] + \gamma LC_2\mathbb{E} [\gamma_k] \leq \|w_{k-1} - w_*\|^2 \\
- \gamma \left( 1 - \gamma L \left( \frac{1}{\alpha_{\text{dwn}}} + \frac{4\omega_{\text{up}}}{N} \right) \gamma LC_2 + \frac{\alpha_{\text{up}}C_1}{N} \right) \mathbb{E} [\langle \nabla F(\bar{w}_{k-1}), \bar{w}_{k-1} - w_* \rangle] \\
+ (2\omega_{\text{up}}(1 + 2\gamma LC_2) + (1 - \alpha_{\text{up}})C_1) \gamma^2\mathbb{E} [\xi_{k-1}] \\
+ \left( 2\gamma L\omega_{\text{dwn}} \left( 1 - \frac{\alpha_{\text{dwn}}}{2} \right) \gamma LC_2 \right) \mathbb{E} [\gamma_{k-1}] \\
+ \frac{\gamma^2 \sigma^2}{N\bar{b}} ((1 + 4\omega_{\text{up}})(1 + 2\gamma LC_2) + 2\alpha_{\text{up}}C_1),
\]

(S29)

We first observe that:

\[
2\gamma L\omega_{\text{dwn}} + \left( 1 - \frac{\alpha_{\text{dwn}}}{2} \right) \gamma LC_2 \leq \gamma LC_2 \iff C_2 \geq \frac{4\omega_{\text{dwn}}}{\alpha_{\text{dwn}}}, \quad \text{which is true by definition of } C_2.
\]

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

\[
2\omega_{\text{up}}(1 + 2\gamma LC_2) + (1 - \alpha_{\text{up}})C_1 \leq C_1 \\
\implies C_1 \geq \frac{2\omega_{\text{up}}(1 + 8\gamma L\omega_{\text{dwn}}/\alpha_{\text{dwn}})}{\alpha_{\text{up}}} \quad \text{because } C_2 = 4\omega_{\text{dwn}}/\alpha_{\text{dwn}}.
\]

Finally, we have that $1 - \gamma L \left( \frac{1}{\alpha_{\text{dwn}}} + \frac{4\omega_{\text{up}}}{N} \right) \gamma LC_2 + \frac{\alpha_{\text{up}}C_1}{N}$ $\geq \frac{1}{2}$, if we take $\gamma$ such that:

\[
\begin{align*}
\left( \frac{1}{\alpha_{\text{dwn}}} + \frac{4\omega_{\text{up}}}{N} \right) (\gamma L)^2 C_2 & \leq 1/4 \implies \gamma \leq \frac{1}{4L \sqrt{\frac{\omega_{\text{dwn}}}{\alpha_{\text{dwn}}} \left( \frac{1}{\alpha_{\text{dwn}}} + \frac{4\omega_{\text{up}}}{N} \right)}} \\
\frac{\gamma LC_2}{N} & \leq 1/4 \iff \frac{2\gamma L\omega_{\text{up}}}{N} (1 + 8\gamma L\omega_{\text{dwn}}/\alpha_{\text{dwn}}) \leq 1/4
\end{align*}
\]

We rewrite the second condition as follows:

\[
\begin{align*}
16(\gamma L)^2 \frac{\omega_{\text{up}}\omega_{\text{dwn}}}{\alpha_{\text{dwn}}N} & \leq 1/8 \iff \gamma \leq \frac{1}{8L \sqrt{\frac{\omega_{\text{dwn}}}{\alpha_{\text{dwn}}} \frac{\omega_{\text{up}}}{N}}} \\
\frac{2\gamma L\omega_{\text{up}}}{N\alpha_{\text{up}}} & \leq 1/8 \iff \gamma \leq \frac{N}{16L\omega_{\text{up}}}.
\end{align*}
\]

Applying convexity, we derive:

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

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

\[
V_k \leq V_{k-1} - \gamma \mathbb{E} [F(w_{k-1}) - F(w_*)] + \frac{\gamma^2 \sigma^2 \Phi^{\text{Heterog}}(\gamma)}{Nb}.
\]
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 \( \alpha_{dwn} \leq \frac{1}{8\omega_{dwn}} \) and \( \alpha_{up}(1 + \omega_{up}) \leq 1 \), for any sequence \((\gamma_k)_{k \in \mathbb{N}} \leq \frac{1}{3\max_{\gamma}}\), for any \( k \in \mathbb{N} \), defining:

\[
V_k := E \left[ \| w_k - w_* \|^2 \right] + \gamma_k^2 C_1 E [\Xi_k] + \gamma_k LC_2 E [\Upsilon_k],
\]

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

\[
V_k \leq (1 - \gamma_k \mu) V_{k-1} - \gamma E [F(\hat{w}_{k-1}) - F(w_*)] + \frac{\gamma^2 \sigma^2 \Phi_{\text{Heterog}}(\gamma)}{Nb}.
\]

**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 \( \gamma_k = 2/(\mu(k + 1) + L) \) that depends on the iteration \( k \) in \( \mathbb{N} \).

We consider this following Lyapunov function:

\[
V_k := E \left[ \| w_k - w_* \|^2 \right] + \gamma_k^2 C_1 E [\Xi_k] + \gamma_k LC_2 E [\Upsilon_k],
\]

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

\[
E \left[ \| w_k - w_* \|^2 \right] + \gamma_k^2 C_1 E [\Xi_k] + \gamma_k LC_2 E [\Upsilon_k] \leq \| w_{k-1} - w_* \|^2
\]

\[
- \gamma_k \left( 1 - \gamma_k L \left( \frac{1 - \alpha_{dwn}}{\alpha_{dwn} N} \right) \gamma_k LC_2 + \frac{\alpha_{up} C_1}{\alpha} \right) \right) E [F(\hat{w}_{k-1}) - F(w_*)]
\]

\[
+ 2\omega_{up}(1 + 2\gamma_k LC_2) + (1 - \alpha_{up} C_1) \gamma_k^2 E [\Xi_{k-1}]
\]

\[
+ 2\gamma_k L\omega_{dwn} + \left( 1 - \frac{\alpha_{dwn}}{2} \right) \gamma_k LC_2 E [\Upsilon_{k-1}]
\]

\[
+ \frac{\gamma_k^2 \sigma^2}{Nb} ((1 + 4\omega_{up})(1 + 2\gamma_k LC_2) + 2\alpha_{up} C_1)
\]

To ensure a \((1 - \gamma \mu)\)-convergence we first choose \( \left( 1 - \alpha_{dwn} C_2 \right)^2 \gamma_k LC_2 \leq (1 - \gamma_k \mu) \gamma_k - 1 \) i.e \( 1 - \gamma_k \mu \leq \frac{\gamma_k}{\gamma_k - 1} \).

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

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

and so, the inequality is always true.

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

Secondly, we need:

\[
(2\omega_{up}(1 + 2\gamma_k LC_2) + (1 - \alpha_{up} C_1) \gamma_k^2 \leq (1 - \gamma_k \mu) \gamma_k - 1 C_1
\]

\[
\iff 2\omega_{up}(1 + 2\gamma_k LC_2) + (1 - \alpha_{up} C_1) \leq \frac{\gamma_k - 1}{\gamma_k} C_1 \quad \text{because} \quad \frac{1 - \gamma_k \mu}{\gamma_k} \leq \frac{1}{\gamma_k - 1},
\]

because \( \gamma_k/\gamma_k \leq \gamma_k - 1/\gamma_k \), it is true if we verify the following stronger condition:

\[
2\omega_{up}(1 + 2\gamma_k LC_2) + (1 - \alpha_{up} C_1) \leq \gamma_k C_1
\]

\[
C_1 \geq \frac{2\omega_{up}(1 + 8\gamma_k L\omega_{dwn}/\alpha_{dwn})}{\alpha_{up}} \quad \text{because} \quad C_2 = 4\omega_{dwn}/\alpha_{dwn}.
\]

57
Finally, in order to apply convexity we must verify:

$$1 - \gamma L \left( \left( \frac{1}{\alpha_{\text{dwn}}} + \frac{4\omega_{\text{up}}}{N} \right) C_2 + \frac{\alpha_{\text{up}} C_1}{N} \right) \geq \frac{1}{2}.$$ 

We take $\gamma_k$ such that:

$$\left\{ \begin{array}{c}
(\frac{1}{\alpha_{\text{dwn}}} + \frac{4\omega_{\text{up}}}{N}) \gamma_k L C_2 \leq \frac{1}{4} \Rightarrow \gamma_k \leq \frac{1}{4L \sqrt{\frac{\omega_{\text{dwn}}}{\alpha_{\text{dwn}}} \left( \frac{1}{\alpha_{\text{dwn}}} + \frac{4\omega_{\text{up}}}{N} \right)}} \\
\frac{\gamma L \alpha_{\text{up}} C_1}{N} \leq \frac{1}{4} \Leftrightarrow \frac{2\gamma_k L \omega_{\text{up}}}{N} (1 + 8\gamma_k L \omega_{\text{dwn}}/\alpha_{\text{dwn}}) \leq \frac{1}{4}
\end{array} \right.$$ 

We rewrite the second condition as following:

$$\left\{ \begin{array}{c}
16(\gamma_k L)^2 \frac{\omega_{\text{dwn}}}{\alpha_{\text{dwn}} N} \leq \frac{1}{8} \Leftrightarrow \gamma_k \leq \frac{1}{8L \sqrt{\frac{\omega_{\text{dwn}}}{\alpha_{\text{dwn}}} \cdot \frac{\omega_{\text{up}}}{N}}} \\
\frac{2\gamma_k L \omega_{\text{up}}}{N} \leq \frac{1}{8} \Leftrightarrow \gamma_k \leq \frac{1}{16L \frac{\omega_{\text{up}}}{N}}
\end{array} \right.$$ 

Now, we can apply strong-convexity:

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

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

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

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

\[\square\]

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]