MCM: Preserved central model for faster bidirectional compression in distributed settings[1]

FLOW: Federated Learning One World Seminar, 29 sept. 2021.

Joint work with **Constantin Philippenko.**

Aymeric DIEULEVEUT
Assistant Professor, École Polytechnique,
Institut Polytechnique de Paris.

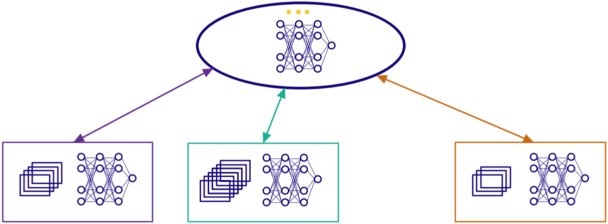
[1] Accepted in Neurips 2021!

Federated Learning and compression

General Framework:

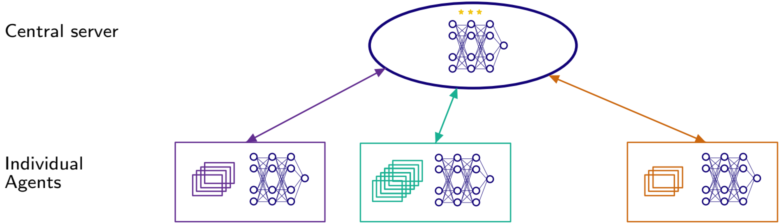
Central server

Individual
Agents



Federated Learning and compression

General Framework:



Constraints

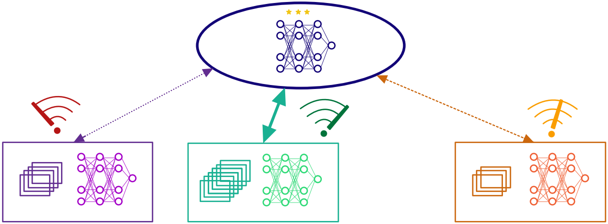
- Heterogeneity

Federated Learning and compression

General Framework:

Central server

Individual Agents



Constraints

- Heterogeneity
- Communication constraints

Compression - multiple directions

Compression is a well identified problem in Federated Learning. [KMA$^+$19].
Multiple very active lines of research:

1. Proposing compression operators.
   - QSGD, Nu-QSGD,
   - Atomo, Power-SGD, HSQ etc. .

Compression - multiple directions

Compression is a well identified problem in Federated Learning. [KMA+19].
Multiple very active lines of research:

1. Proposing compression operators.
   - QSGD, Nu-QSGD,
   - Atomo, Power-SGD, HSQ etc. .
2. Studying the impact of the properties of algorithms on convergence:
   - Biased vs Unbiased
   - Independent or not
   - Bounded variance, relatively bounded variance,
   - Adaptation

Compression - multiple directions

Compression is a well identified problem in Federated Learning. [KMA+19].
Multiple very active lines of research:

1. Proposing compression operators.
   - QSGD, Nu-QSGD,
   - Atomo, Power-SGD, HSQ etc. .
2. Studying the impact of the properties of algorithms on convergence:
   - Biased vs Unbiased
   - Independent or not
   - Bounded variance, relatively bounded variance,
   - Adaptation
3. Adapting algorithms with compression.
   Even if we communicate at each step, compression can prevent the algorithm from converging.
   - **Impact of Bias in the compression operator. Error-Feedback** line of work [SCJ18, SK19]
   - **Impact of Heterogeneity**. **Memory** line of work [MGTR19].

Outline

1. **Part 1: Preserved iterate for double compression in distributed-heterogeneous framework.**
   ↬ Adapting algorithms with compression
   Joint work with Constantin Philippenko

2. Another time: **DoStoVoQ** ↬ Proposing compression operators, Studying the impact of the properties of algorithms on convergence,

Bi-directional compression

To limit the number of bits exchanged, we **compress** each signal before transmitting it. We introduce compression operators $\mathcal{C}_{\mathrm{dwn}}$ and $\mathcal{C}_{\mathrm{up}}$.

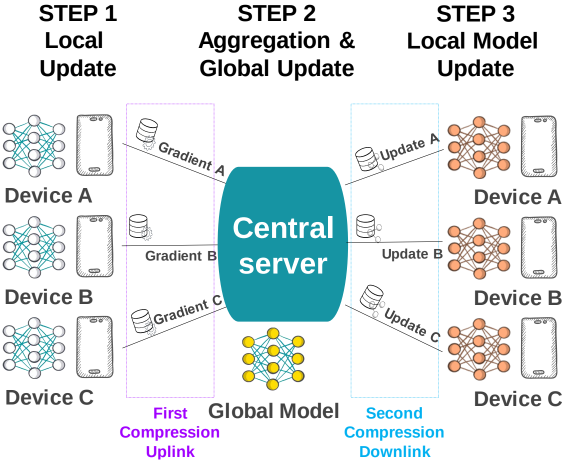


Figure: Bidirectional compression. 1) Uplink: compress the gradients. 2) Downlink: compress the update.

1. Bi-directional compression

We introduce compression operators $\mathcal{C}_{\mathrm{dwn}}$ and $\mathcal{C}_{\mathrm{up}}$.

Assumption 1

For $\mathrm{dir} \in \{\mathrm{up}, \mathrm{dwn}\}$, *there exists a constant* $\omega_{\mathrm{dir}} \in \mathbb{R}^*$ *s.t.* $\mathcal{C}_{\mathrm{dir}}$ *satisfies. for all* Δ *in* $\mathbb{R}^d$ *:*

$$\mathbb{E}[\mathcal{C}_{\mathrm{dir}}(\Delta)] = \Delta \quad \textit{and} \quad \mathbb{E}\left[\|\mathcal{C}_{\mathrm{dir}}(\Delta) - \Delta\|^2\right] \leq \omega_{\mathrm{dir}} \|\Delta\|^2 .$$

Several well-known compression operator: quantization, sparsification, etc. .

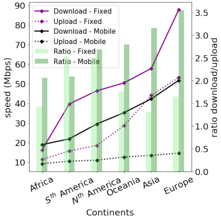
Do we need double compression?

Objectives of compression:

1. Accelerate the learning process,
2. Limit the number of communicated bits

In terms of speed, double compression depends on how the exchange is performed:

- If broadcast (1 to N) is much faster than upload (N to 1) then no need for double compression.
- if we consider mobile devices (using for example fast Internet connexion), only a small difference between upload and download speed.



In terms of communicated bits: upload and download are symmetric.
Example: no one wants to download a large update everyday on its phone!

Double compression: first attempts and related work

⇒ The update equation becomes: $w_k = w_{k-1} - \gamma \mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\mathrm{up}}(g_k^i) \right)$

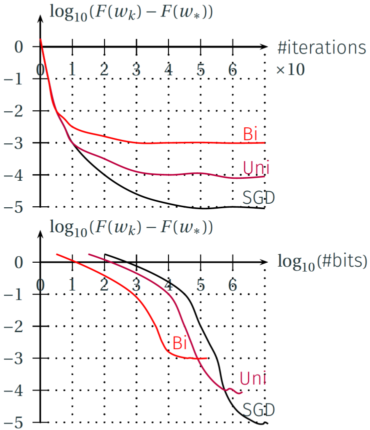
Table: Features of the main existing algorithms performing compression. e_k^i (resp. E_k) denotes the use of error-feedback at uplink (resp. downlink). h_k^i (resp. H_k) denotes the use of a memory at uplink (resp. downlink). Note that `Dist-EF-SGD` is identical to `Double-Squeeze` but has been developed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

	Compr.	e_k^i	h_k^i	E_k	H_k	Rand.	update point
`Qsgd` [AGL+17]	one-way						
`ECQ-sgd` [WHHZ18]	one-way	✓					
`Diana` [MGTR19]	one-way		✓				
`Dore` [LLTY20]	two-way		✓	✓			degraded
`Double-Squeeze` [TYL+19], `Dist-EF-SGD` [ZHK19]	two-way	✓		✓			degraded
`Artemis` [PD20]	two-way		✓				degraded
Doubly compressed SGD [GKMR20]	two-way		✓				degraded
`MCM`	two-way		✓		✓		non-degraded
`Rand-MCM`	two-way		✓		✓	✓	non-degraded

Precise comparison of convergence results will be given afterwards.

Expected results for Double compression

1. The level of noise in the gradient increases,
2. Proportionally to ω_{dwn}
3. In fact, we can prove that the limit Variance indeed provably increases [PD20].



2. The memory mechanism

Motivation: The distribution of the observations on worker i and j are often different.

Assumption 2

For all $i \in [N]$:

$$\|\nabla F_i(w_*)\|^2 \leq B^2$$

Challenge: Compression of a quantity that goes to 0 !

2. The memory mechanism

Motivation: The distribution of the observations on worker i and j are often different.

Assumption 2

For all $i \in [N]$:

$$\|\nabla F_i(w_*)\|^2 \leq B^2$$

Challenge: Compression of a quantity that goes to 0 !
Solution: Compute (on the server and the worker independently) a **"memory"** h_k^i s.t. $h_k^i \to_{k\to\infty} \nabla F_i(w_*)$.

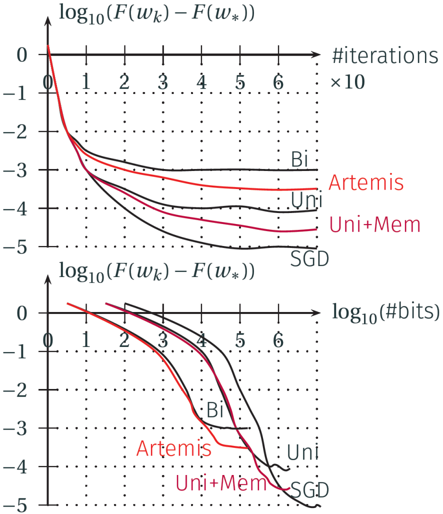
$\Rightarrow$ The update equation becomes:

$$w_k = w_{k-1} - \gamma \mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\mathrm{up}}(g_k^i - h_k^i) + h_k^i \right)$$
$$h_{k+1}^i = h_k^i + \alpha \mathcal{C}_{\mathrm{up}}(g_k^i - h_k^i)$$

Crucial role of (uplink)-memory on heterogeneous data. [MGTR19, PD20].

The memory mechanism

Expected improvement with uplink memory in the heterogeneous framework.



The non-degraded update

Classical double compression (e.g., `Artemis`)- **compress the update sent back to the workers and use it to update the model.**

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

The gradient is taken at the point w_k held by the central server.

The non-degraded update

Classical double compression (e.g., `Artemis`)- **compress the update sent back to the workers and use it to update the model.**

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

The gradient is taken at the point w_k held by the central server.

MCM - **preserve the model on the central server.**

$$\begin{aligned} w_k &= w_{k-1} - \gamma \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\mathrm{up}}(g_k^i(\hat{w}_{k-1})) \right) \\ \hat{w}_k &= w_{k-1} - \gamma \mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\mathrm{up}}(g_k^i(\hat{w}_{k-1})) \right) \end{aligned} \quad (1)$$

The gradient is taken at a random point $\hat{w}_k$ s.t. $\mathbb{E}[\hat{w}_k | w_k] = w_k$

The non-degraded update

Classical double compression (e.g., `Artemis`)- **compress the update sent back to the workers and use it to update the model.**

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

The gradient is taken at the point w_k held by the central server.

MCM - **preserve the model on the central server.**

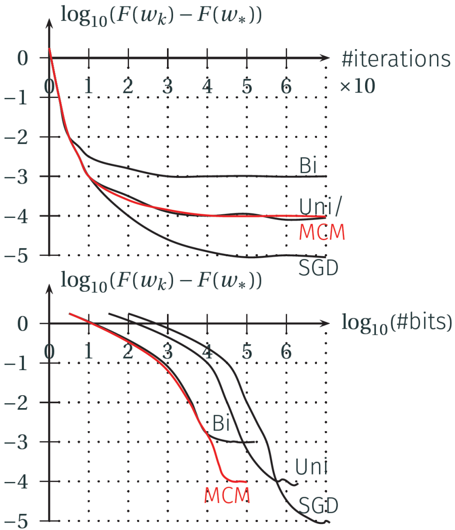
$$\begin{aligned} w_k &= w_{k-1} - \gamma \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\mathrm{up}}(g_k^i(\hat{w}_{k-1})) \right) \\ \hat{w}_k &= w_{k-1} - \gamma \mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\mathrm{up}}(g_k^i(\hat{w}_{k-1})) \right) \end{aligned} \tag{1}$$

The gradient is taken at a random point $\hat{w}_k$ s.t. $\mathbb{E}[\hat{w}_k | w_k] = w_k$

Update (1) is not feasible in practice. We refer to this algorithm as a `Ghost` algorithm.

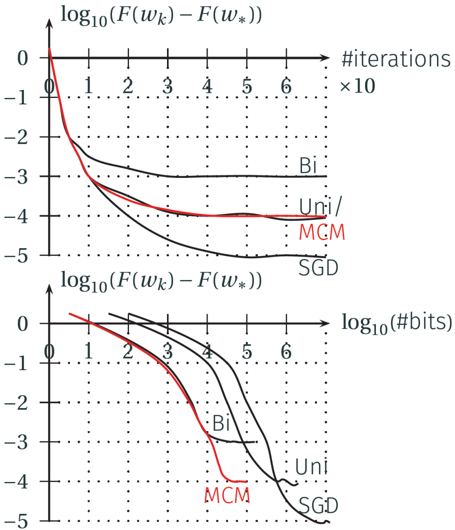
Ghost algorithm

What do we hope for?



Ghost algorithm

What do we hope for?



Outline towards proof of convergence:

1. Assumptions
2. Convergence of Ghost
3. Sketch of proof
4. Adaptation into a practical algorithm
5. Extensions

Assumptions

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

Assumption 3 (Smoothness)

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

Assumptions

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

Assumption 3 (Smoothness)

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

Assumption 4 (Convexity)

F is convex, that is for all vectors w_1, w_2 in $\mathbb{R}^d$: $F(w_2) \geq F(w_1) + (w_2 - w_1)^T \nabla F(w_1)$.

Assumptions

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

Assumption 3 (Smoothness)

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

Assumption 4 (Convexity)

F is convex, that is for all vectors w_1, w_2 in $\mathbb{R}^d$: $F(w_2) \geq F(w_1) + (w_2 - w_1)^T \nabla F(w_1)$.

Assumption 5 (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: $E[\|g_k^i(w) - \nabla F(w)\|^2] \leq \sigma^2/b$.

Convergence of Ghost

Definition 1 (Ghost algorithm)

Recall that the Ghost algorithm is defined as follows, for $k \in \mathbb{N}$, for all $i \in [\![1, N]\!]$ we have:

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

Proposition 1

Consider the Ghost update in eq. (1)*, under Assumptions 1, 3 and 5, for all k in $\mathbb{N}$ with the convention $\nabla F(w_{-1}) = 0$:*

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

Sketch of Proof

Proof.

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

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

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

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

then we do a Bias Variance decomposition. □

↬ the variance of the local model is bounded by an affine function of the squared norm of the *previous* stochastic gradients $\nabla F(\widehat{w}_{k-1})$.

Sketch of proof, 2

Then, classical perturbed iterate approach [MPP+16],

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

Moreover,

$$\begin{aligned} -2\gamma \mathbb{E}\left\langle \nabla F(\widehat{w}_{k-1}) \mid w_{k-1} - w_* \right\rangle &= -2\gamma \mathbb{E}\left\langle \nabla F(\widehat{w}_{k-1}) \mid \widehat{w}_{k-1} - w_* \right\rangle \\ &\quad + 2\gamma \mathbb{E}\left\langle \nabla F(\widehat{w}_{k-1}) - \nabla F(w_{k-1}) \mid w_{k-1} - \widehat{w}_{k-1} \right\rangle. \end{aligned}$$

as $\mathbb{E}\left[\widehat{w}_{k-1} \mid w_{k-1}\right] = w_{k-1}$.

Sketch of proof, 2

Then, classical perturbed iterate approach [MPP+16],

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

Moreover,

$$-2\gamma \mathbb{E}\left\langle \nabla F(\widehat{w}_{k-1}) \mid w_{k-1} - w_* \right\rangle = -2\gamma \mathbb{E}\left\langle \nabla F(\widehat{w}_{k-1}) \mid \widehat{w}_{k-1} - w_* \right\rangle + 2\gamma \mathbb{E}\left\langle \nabla F(\widehat{w}_{k-1}) - \nabla F(w_{k-1}) \mid w_{k-1} - \widehat{w}_{k-1} \right\rangle.$$

as $\mathbb{E}[\widehat{w}_{k-1} \mid w_{k-1}] = w_{k-1}$.

1. $-2\gamma \mathbb{E}\left\langle \nabla F(\widehat{w}_{k-1}) \mid \widehat{w}_{k-1} - w_* \right\rangle$ "strong contraction", upper bounded by
   - $-2\gamma(\mu\|\widehat{w}_{k-1} - w_*\|^2 + F(\widehat{w}_{k-1}) - F_*)$
   - $-2\gamma\|\nabla F(\widehat{w}_{k-1})\|^2 / L$
2. $2\gamma \mathbb{E}\left\langle \nabla F(\widehat{w}_{k-1}) - \nabla F(w_{k-1}) \mid w_{k-1} - \widehat{w}_{k-1} \right\rangle$ positive residual term.

Theorem 2 (Contraction for Ghost, convex case)

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

Contraction for Ghost

Theorem 3 (Contraction for Ghost, convex case)

Under Assumptions 1 and 3 to 5, with $\mu = 0$*, if* $\gamma L(1 + \omega_{\mathrm{up}}/N) \leq \frac{1}{2}$.

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

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 $\hat{w}_{k-2}$.
2. if $2\gamma^3 \omega_{\mathrm{dwn}} L(1 + \omega_{\mathrm{up}}/N) \leq \gamma/(2L)$, then these terms eventually cancel out.
3. This is equivalent to $2\gamma L\sqrt{\omega_{\mathrm{dwn}}(1 + \omega_{\mathrm{up}}/N)} \leq 1$. It is natural to chose $\gamma \leq 1/(2L\max(1 + \omega_{\mathrm{up}}/N, 1 + \omega_{\mathrm{dwn}}))$.

Line of proof is the same for strongly convex, but different for non-convex.

Noise level, Ghost

Theorem 4 (Contraction for Ghost, convex case)

Under Assumptions 1 and 3 to 5, with $\mu = 0$, if $\gamma L(1 + \omega_{\mathrm{up}}/N) \leq \frac{1}{2}$.

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

For Ghost algorithm

$$\gamma^2\frac{(1+\omega_{\mathrm{up}})\sigma^2}{Nb}(1 + 2\gamma L\omega_{\mathrm{dwn}}).$$

For classical double compression

$$\gamma^2\frac{\omega_{\mathrm{dwn}}(1+\omega_{\mathrm{up}})\sigma^2}{Nb}.$$

For unidirectional-compression

$$\gamma^2\frac{(1+\omega_{\mathrm{up}})\sigma^2}{Nb}.$$

A practical algorithm?

Summary:

1. For a hypothetical iterate, we can obtain convergence in the “preserved central iterate” framework
2. The limit Variance is nearly of the same order as with simple compression.
3. This algorithm cannot be implemented in practice!

A practical algorithm?

Summary:

1. For a hypothetical iterate, we can obtain convergence in the "preserved central iterate" framework
2. The limit Variance is nearly of the same order as with simple compression.
3. This algorithm cannot be implemented in practice!

New attempts:

Ghost

$$w_k = w_{k-1} - \gamma \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\mathrm{up}}(g_k^i(\hat{w}_{k-1})) \right)$$

$$\hat{w}_k = w_{k-1} - \gamma \mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\mathrm{up}}(g_k^i(\hat{w}_{k-1})) \right)$$

Update compression

$$w_k = w_{k-1} - \gamma \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\mathrm{up}}(g_k^i(\hat{w}_{k-1})) \right)$$

$$\hat{w}_k = \hat{w}_{k-1} - \gamma \mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\mathrm{up}}(g_k^i(\hat{w}_{k-1})) \right)$$

Model compression ($\alpha_{\mathrm{dwn}} = 0$)

$$w_k = w_{k-1} - \gamma \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\mathrm{up}}(g_k^i(\hat{w}_{k-1})) \right)$$

$$\hat{w}_k = \mathcal{C}_{\mathrm{dwn}}(w_k)$$

Model difference compression ($\alpha_{\mathrm{dwn}} = 1$)

$$w_k = w_{k-1} - \gamma \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\mathrm{up}}(g_k^i(\hat{w}_{k-1})) \right)$$

$$\hat{w}_k = \hat{w}_{k-1} - \gamma \mathcal{C}_{\mathrm{dwn}}(w_k - \hat{w}_{k-1})$$

First attempts - Variance of the local iterate is too high.

- Update compression
- Model difference compression ($\alpha_{\mathrm{dwn}} = 1$)
- Model compression ($\alpha_{\mathrm{dwn}} = 0$)
- MCM

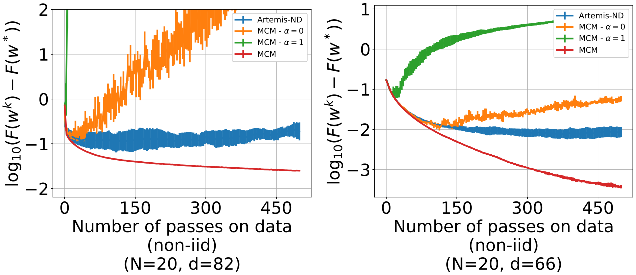


Figure: Comparing MCM on two datasets with three other algorithms using a non-degraded update, $\gamma = 1/L$.

The downlink memory mechanism for MCM

We introduce a *downlink memory term* $(H_k)_{k\in\mathbb{N}}$:

1. available on both workers and central server
2. the difference Ω_{k+1} between the model and this memory is compressed and exchanged
3. the local model is reconstructed from this information

$$\begin{cases} w_k & = w_{k-1} - \gamma \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\mathrm{up}}(g_k^i(\hat{w}_{k-1})) \right) \\ \Omega_{k+1} & = w_{k+1} - H_k \\ \widehat{w}_{k+1} & = H_k + \mathcal{C}_{\mathrm{dwn}}(\Omega_{k+1}) \\ H_{k+1} & = H_k + \alpha_{\mathrm{dwn}} \mathcal{C}_{\mathrm{dwn}}(\Omega_{k+1}). \end{cases} \tag{3}$$

Introducing this memory mechanism is crucial to control the variance of the local model $\widehat{w}_{k+1}$.

Control of the local Variance

Let $\Upsilon_k := \|w_k - H_{k-1}\|^2$.

Theorem 5

Consider the MCM update. Under Assumptions 1, 3 and 5 with $\mu = 0$, if $\gamma \leq (8\omega_{\mathrm{dwn}} L)^{-1}$ and $\alpha_{\mathrm{dwn}} \leq (4\omega_{\mathrm{dwn}})-1$, then for all k in $\mathbb{N}$:

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

Convergence of MCM - **Convex**

Let

1. $V_k = \mathbb{E}[\|w_k - w_*\|^2] + 32\gamma L\omega_{\mathrm{dwn}}^2\mathbb{E}[\Upsilon_k]$.
2. $\Phi(\gamma) := (1+\omega_{\mathrm{up}})\left(1+64\gamma L\omega_{\mathrm{dwn}}^2\right)$.

Convergence of MCM - **Convex**

Let

1. $V_k = \mathbb{E}[\|w_k - w_*\|^2] + 32\gamma L \omega_{\mathrm{dwn}}^2 \mathbb{E}[\Upsilon_k]$.
2. $\Phi(\gamma) := (1 + \omega_{\mathrm{up}}) \left(1 + 64\gamma L \omega_{\mathrm{dwn}}^2\right)$.

Theorem 6 (Convergence of MCM, convex case)

Under Assumptions 1 and 3 to 5 with $\mu = 0$. *For all* $k > 0$, *for any* $\gamma \leq \gamma_{\max}$, *we have, for* $\bar{w}_k = \frac{1}{k}\sum_{i=0}^{k-1} w_i$,

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

Convergence of MCM - **Convex**

Let

1. $V_k = \mathbb{E}[\|w_k - w_*\|^2] + 32\gamma L\omega_{\mathrm{dwn}}^2 \mathbb{E}[\Upsilon_k]$.
2. $\Phi(\gamma) := (1 + \omega_{\mathrm{up}}) \left(1 + 64\gamma L\omega_{\mathrm{dwn}}^2\right)$.

Theorem 6 (Convergence of MCM, convex case)

Under Assumptions 1 and 3 to 5 with $\mu = 0$. For all $k > 0$, for any $\gamma \leq \gamma_{\max}$, we have, for $\bar{w}_k = \frac{1}{k}\sum_{i=0}^{k-1} w_i$,

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

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

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

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

Comparison to previous results: **Limit Variance**

Better limit variance $\Rightarrow$ better rate.

For a constant γ,

1. the variance term is upper bounded by
$$\frac{\gamma^2\sigma^2}{Nb}(1+\omega_{\mathrm{up}})(1+64\gamma L\omega_{\mathrm{dwn}}^2).$$
2. impact of the downlink compression is attenuated by a factor γ. As $\gamma \to 0$ we get close to `Diana`, i.e., without downlink compression [MGTR19, Eq. 16 in Th. 2]
$$\frac{\gamma^2\sigma^2}{Nb}(1+\omega_{\mathrm{up}}).$$
3. This is much lower than the variance for previous algorithms using double compression for
$$\gamma^2\sigma^2(1+\omega_{\mathrm{up}})(1+\omega_{\mathrm{dwn}})/N$$

   - for `Dore`, see Corollary 1 in Liu et al. [LLTY20] (who indicate $(1-\rho)^{-1} \geq (1+\omega_{\mathrm{up}}/N)(1+\omega_{\mathrm{dwn}})$),
   - for `Artemis` see Table 2 and Th. 3 point 2 in [PD20],
   - for Gorbunov et al. [GKMR20], see Theorem I.1. (with $\gamma D_1' \propto \gamma^2\sigma^2(1+\omega_{\mathrm{up}})(1+\omega_{\mathrm{dwn}})/N$).

Comparison to previous results: **Limit learning rate**

Limit learning rate: Maximal learning rate to ensure convergence.
$\gamma_{\max} := \min(\gamma_{\max}^{\mathrm{up}}, \gamma_{\max}^{\mathrm{dwn}}, \gamma_{\max}^{\Upsilon})$, where

- $\gamma_{\max}^{\mathrm{up}} := (2L(1+\omega_{\mathrm{up}}/N))^{-1}$ corresponds to the classical constraint on the learning rate in the unidirectional regime [see MGTR19, PD20],
- $\gamma_{\max}^{\mathrm{dwn}} := (8L\omega_{\mathrm{dwn}})^{-1}$ is a similar constraint coming from the downlink compression,
- $\gamma_{\max}^{\Upsilon} := \left(8\sqrt{2}L\omega_{\mathrm{dwn}}\sqrt{8\omega_{\mathrm{dwn}}+\omega_{\mathrm{up}}/N}\right)^{-1}$ is a combined constraint that arises when controlling the variance term Υ.[2]

Remarks

- weaker constraints than in the “degraded” framework [LLTY20, PD20], in which $\gamma_{\max}^{\mathrm{Dore}} \leq \left(8L(1+\omega_{\mathrm{dwn}})(1+\omega_{\mathrm{up}}/N)\right)^{-1}$.
- e.g., if $\omega_{\mathrm{up,dwn}} \to \infty$ and $\omega_{\mathrm{dwn}} \simeq \omega_{\mathrm{up}} \simeq: \omega$, the maximal learning rate for MCM is $(L\omega^{3/2})^{-1}$, while it is $(L\omega^2)^{-1}$ in [LLTY20, PD20]. Our $\gamma_{\max}$ is thus larger by a factor $\sqrt{\omega}$.

[2] The dependency in $\omega^{3/2}$ is similar to the one obtained by Horváth [HKM+19] in unidirectional compression in the non-convex case (Theorem 4).

Convergence of MCM - **Strongly Convex**

We define $\widetilde{L}$ such that $\gamma_{\max} = (2\widetilde{L})^{-1}$.

Theorem 7 (Convergence of MCM in the homogeneous and strongly-convex case)

Under Assumptions 1 and 3 to 5 with $\mu > 0$, for k in $\mathbb{N}$, for any sequence $(\gamma_k)_{k\geq 0} \leq \gamma_{\max}$:

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

where $\Phi(\gamma_k) = (1 + \omega_{\mathrm{up}})\left(1 + 64\gamma_k L \omega_{\mathrm{dwn}}^2\right)$. Consequently,

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

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

Summary of rates and complexities

Summary of rates. In this Table, 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 ω_{up}, ω_{dwn}, and complexities for $\epsilon \to 0$.

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

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

Extensions - and partial take away

1. Heterogeneous framework: previous theorems are valid in the heterogeneous framework (at the cost of a constant 2), under Assumption 2.
2. Another theorem is provided in the non-convex regime, with similar take-away.

Take away:

1. MCM= Model Compression with memory
2. Uses a **memory** on the downlink direction, as introduced by Mishchenko [MGTR19] for the uplink.
3. Leverages the unbiased-ness of $\hat{w}_k$ around w_k.

Next step: worker dependent downlink compression: Rand-MCM!

No (or few) reasons to use the same compression for all workers !

$$\begin{cases} w_k &= w_{k-1} - \gamma \left(\frac{1}{N} \sum_{i=1}^N \mathcal{C}_{\mathrm{up}}(g_k^i(\hat{w}_{k-1}^i)) \right) \\ \Omega_{k+1} &= w_{k+1} - H_k \\ \widehat{w}_{k+1}^i &= H_k{}^i + \mathcal{C}_{\mathrm{dwn},i}(\Omega_{k+1}) \\ H_{k+1}^i &= H_k{}^i + \alpha_{\mathrm{dwn}} \mathcal{C}_{\mathrm{dwn},i}(\Omega_{k+1}). \end{cases} \tag{4}$$

Advantages:

1. Independence could help reduce the variance
2. Workers can be allowed to choose the size (or equivalently the compression level) of their updates.
3. Helps in case of Partial Participation
4. Could be leveraged to tackle *honest-but-curious clients*.

Next step: worker dependent downlink compression: Rand-MCM!

No (or few) reasons to use the same compression for all workers !

$$\begin{cases} w_k &= w_{k-1} - \gamma \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\mathrm{up}}(g_k^i(\hat{w}_{k-1}^i)) \right) \\ \Omega_{k+1} &= w_{k+1} - H_k \\ \widehat{w}_{k+1}^i &= H_k{}^i + \mathcal{C}_{\mathrm{dwn},i}(\Omega_{k+1}) \\ H_{k+1}^i &= H_k{}^i + \alpha_{\mathrm{dwn}} \mathcal{C}_{\mathrm{dwn},i}(\Omega_{k+1}). \end{cases} \tag{4}$$

Advantages:

1. Independence could help reduce the variance
2. Workers can be allowed to choose the size (or equivalently the compression level) of their updates.
3. Helps in case of Partial Participation
4. Could be leveraged to tackle *honest-but-curious clients*.

Drawbacks

1. Storing the N memories $(H_k^i)_{i \in [N]}$ instead of one

Next step: worker dependent downlink compression: Rand-MCM!

No (or few) reasons to use the same compression for all workers !

$$\begin{cases} w_k & = w_{k-1} - \gamma \left(\frac{1}{N} \sum_{i=1}^N \mathcal{C}_{\mathrm{up}}(g_k^i(\hat{w}_{k-1}^i)) \right) \\ \Omega_{k+1} & = w_{k+1} - H_k \\ \widehat{w}_{k+1}^i & = H_k{}^i + \mathcal{C}_{\mathrm{dwn},i}(\Omega_{k+1}) \\ H_{k+1}^i & = H_k{}^i + \alpha_{\mathrm{dwn}} \mathcal{C}_{\mathrm{dwn},i}(\Omega_{k+1}). \end{cases} \tag{4}$$

Advantages:

1. Independence could help reduce the variance
2. Workers can be allowed to choose the size (or equivalently the compression level) of their updates.
3. Helps in case of Partial Participation
4. Could be leveraged to tackle *honest-but-curious clients*.

Drawbacks

1. Storing the N memories $(H_k^i)_{i \in [N]}$ instead of one

Solutions:

1. Keep and use a single memory $\bar{H}_k = N^{-1} \sum_{i=1}^N H_k^i$.
   - 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)
2. Use Rand-MCM with an arbitrary number of groups $G \ll N$ of workers. In each group $\mathcal{G}_g$, $g \in [G]$, all workers share the same memory (H_k^g) and receive the same update $\mathcal{C}_{\mathrm{dwn},g}(w_{k+1} - H_k^g)$. We call this algorithm Rand-MCM-G.

Convergence of Rand-MCM

1. At least as good:

Theorem 8

Theorems 5 to 7 are valid for `Rand-MCM` *and* `Rand-MCM-G`*.*

Convergence of Rand-MCM

1. At least as good:

Theorem 8

Theorems 5 to 7 are valid for `Rand-MCM` *and* `Rand-MCM-G`*.*

2. Better on residual term:

Theorem 9 (Convergence in the quadratic case)

Under Assumptions 1 and 3 to 5 with $\mu = 0$, if the function is quadratic, after running $K > 0$ iterations, for any $\gamma \leq \gamma_{\max}$, and we have

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

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

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 [Bac10]) to control of $\mathbb{E}\left[\|\nabla F(\widehat{w}_{k-1}) - \mathbb{E}[\nabla F(\widehat{w}_{k-1})]\|^2 \;\middle|\; w_{k-1} \right]$.

Experiments

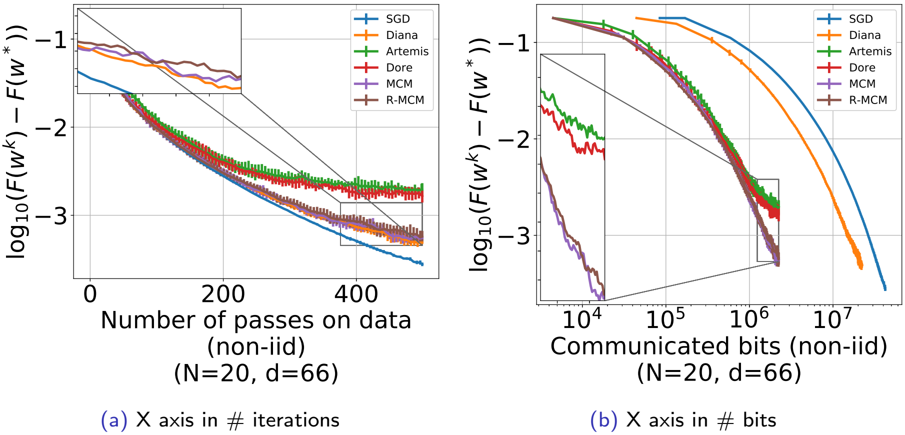
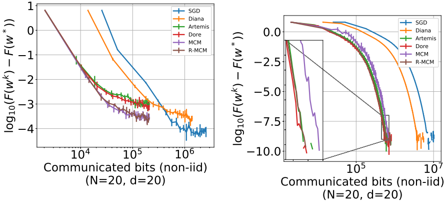


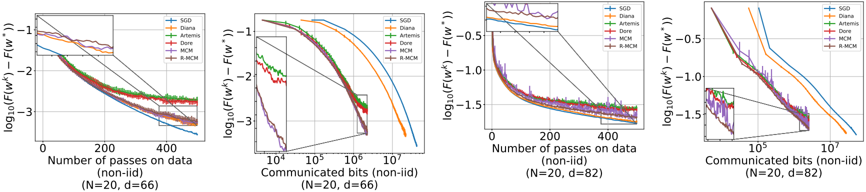
Figure: Quantum with $b = 400$, $\gamma = 1/L$ (LSR).

More experiments



(a) $\sigma^2 \neq 0$, $\gamma = (L\sqrt{k})^{-1}$ (b) $\sigma^2 = 0$, $\gamma = L^{-1}$ (c) MNIST with a CNN (d) CIFAR10 with LeNet

Figure: Convergence on toy dataset on LSR (a,b) and on neural networks (c, d).



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

More experiments (convex)

Excess loss after **450** epochs	SGD	DIANA	MCM	DORE
a9a b=50	-3.5	-2.7	-2.7	-1.8
Phishing b=50	-3.7	-3.5	-3.4	-2.7
w8a b=8	-3.5	-3.0	-2.5	-1.75
Compression	no	uni-dir	bi-dir	bi-dir

More experiments, non convex

Nonconvex framework	MNIST (CNN, d=2e4, 2 bits-quantization with norm 2)	Fashion MNIST (FashionSimpleNet, d=4e5, 2 bits-quantization with norm 2)	Heterogeneous EMNIST (CNN, d=2e4, 2 bits-quantization with norm 2)	CIFAR-10 (LeNet, d=62e3, 2-bits-quantization with norm inf)
Baseline accuracy for the selected network [Ref]		92.3% [Link]		67.52% [Link]
Accuracy after 300 epochs	SGD: 99.0%	SGD: 92.4%	SGD: 99.0%	SGD: 69.1%
	Diana: 98.9% MCM: 98.8% Artemis: 97.9% Dore: 97.9%	Diana: 92.4% MCM: 90.6% Artemis: 86.7% Dore: 87.9%	Diana: 98.9% MCM: 98.9% Artemis: 98.3% Dore: 98.5%	Diana: 64.0% MCM: 63.5% Artemis: 54.8% Dore: 56.3%
Train loss after 300 epochs	SGD:0.025	SGD: 0.093	SGD: 0.026	SGD: 0.909
	Diana: 0.034 MCM: 0.033 Artemis: 0.075 Dore: 0.072	Diana: 0.141 MCM: 0.209 Artemis: 0.332 Dore: 0.300	Diana: 0.031 MCM: 0.030 Artemis: 0.052 Dore: 0.048	Diana: 1.047 MCM: 1.096 Artemis: 1.342 Dore: 1.292

Experiments: Randomization + single memory.

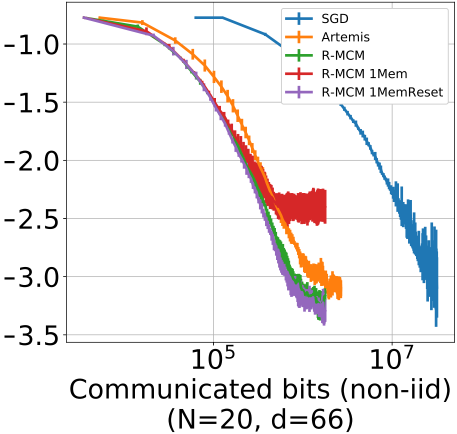


Figure: `Rand-MCM` (PP) on *quantum* with a *single memory* ($s = 2$).

Conclusion and open directions

MCM underlines the importance to not degrade the global model.

Summary:

2. New algorithm for bi-directional compression with a preserved central model
3. Reduces (nearly cancels) impact of downlink compression
4. Achieves the same asymptotic rate of convergence as unidirectional compression.

Conclusion and open directions

MCM underlines the importance to not degrade the global model.

Summary:

2. New algorithm for bi-directional compression with a preserved central model
3. Reduces (nearly cancels) impact of downlink compression
4. Achieves the same asymptotic rate of convergence as unidirectional compression.

Open directions:

1. Can we provably benefit from the smoothing effect?
2. Extending proofs of `Rand-MCM` to the self-concordant framework
3. Leveraging the randomization effect in applications
4. Even better double compression:
    - combination with better techniques on the up-link direction
    - unaffected $\gamma_{\max}$
    - biased compression operators.

Conclusion and open directions

MCM underlines the importance to not degrade the global model.

Summary:

2. New algorithm for bi-directional compression with a preserved central model
3. Reduces (nearly cancels) impact of downlink compression
4. Achieves the same asymptotic rate of convergence as unidirectional compression.

Open directions:

1. Can we provably benefit from the smoothing effect?
2. Extending proofs of `Rand-MCM` to the self-concordant framework
3. Leveraging the randomization effect in applications
4. Even better double compression:
   - combination with better techniques on the up-link direction
   - unaffected $\gamma_{\max}$
   - biased compression operators.

Thank you for your attention :)

Advertisement

1. I am looking for excellent students and postdocs to work on various aspects of Federated Learning in Paris!
2. Research visits can also be organized (3 month+)

[AGL+17] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding. Advances in Neural Information Processing Systems, 30:1709–1720, 2017.

[Bac10] Francis Bach. Self-concordant analysis for logistic regression. Electronic Journal of Statistics, 4(none):384–414, January 2010. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

[GKMR20] Eduard Gorbunov, Dmitry Kovalev, Dmitry Makarenko, and Peter Richtárik. Linearly Converging Error Compensated SGD. arXiv:2010.12292 [cs, math], October 2020. arXiv: 2010.12292.

[HKM+19] Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic Distributed Learning with Gradient Quantization and Variance Reduction. arXiv:1904.05115 [math], April 2019. arXiv: 1904.05115.

[KMA+19] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and Open Problems in Federated Learning. arXiv:1912.04977 [cs, stat], December 2019. arXiv: 1912.04977.

[LLTY20] Xiaorui Liu, Yao Li, Jiliang Tang, and Ming Yan. A Double Residual Compression Algorithm for Efficient Distributed Learning. In International Conference on Artificial Intelligence and Statistics, pages 133–143, June 2020. ISSN: 1938-7228 Section: Machine Learning.

[MGTR19] Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed Learning with Compressed Gradient Differences. arXiv:1901.09269 [cs, math, stat], June 2019. arXiv: 1901.09269.

[MPP+16] Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I. Jordan. Perturbed Iterate Analysis for Asynchronous Stochastic Optimization. arXiv:1507.06970 [cs, math, stat], March 2016. arXiv: 1507.06970.

[PD20] Constantin Philippenko and Aymeric Dieuleveut. Artemis: tight convergence guarantees for bidirectional compression in Federated Learning. arXiv:2006.14591 [cs, stat], November 2020. arXiv: 2006.14591.

[SCJ18] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with Memory. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems 31, pages 4447–4458. Curran Associates, Inc., 2018.

[SFD+14] F. Seide, H. Fu, Jasha Droppo, G. Li, and D. Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. pages 1058–1062, January 2014.

[SK19] Sebastian U. Stich and Sai Praneeth Karimireddy. The Error-Feedback Framework: Better Rates for SGD with Delayed Gradients and Compressed Communication. arXiv:1909.05350 [cs, math, stat], September 2019. arXiv: 1909.05350.

[TYL+19] Hanlin Tang, Chen Yu, Xiangru Lian, Tong Zhang, and Ji Liu. DoubleSqueeze: Parallel Stochastic Gradient Descent with Double-pass Error-Compensated Compression. In International Conference on Machine Learning, pages 6155–6165. PMLR, May 2019. ISSN: 2640-3498.

[WHHZ18] Jiaxiang Wu, Weidong Huang, Junzhou Huang, and Tong Zhang. Error Compensated Quantized SGD and its Applications to Large-scale Distributed Optimization. In International Conference on Machine Learning, pages 5325–5333. PMLR, July 2018. ISSN: 2640-3498.

[ZHK19] Shuai Zheng, Ziyue Huang, and James Kwok. Communication-Efficient Distributed Blockwise Momentum SGD with Error-Feedback. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pages 11450–11460. Curran Associates, Inc., 2019.