Bi-directional compression for Federated Learning: `Artemis` & `MCM`

Aymeric Dieuleveut
CMAP, École Polytechnique, Institut Polytechnique de Paris

Joint work with Constantin Philippenko

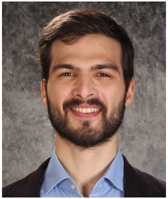


Table of contents

General Federated Learning framework

`Artemis`: a framework for bi-compression in heterogeneous settings

- Theorems
- Experiments

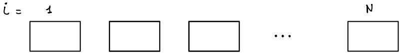
Reducing the impact of downlink compression: `MCM`

General Federated Learning framework

General Federated Learning framework

Learning from a set of N agents: $\min_{w \in \mathbb{R}^d} \left\{ F(w) := \frac{1}{N} \sum_{i=1}^{N} \underbrace{\mathbb{E}_{z \sim \mathscr{D}_i} [\ell(z, w)]}_{F_i(w)} \right\}.$

Learning from a set of N agents: $\min_{w \in \mathbb{R}^d} \left\{ F(w) := \frac{1}{N} \sum_{i=1}^{N} \underbrace{\mathbb{E}_{z \sim \mathcal{D}_i} [\ell(z, w)]}_{F_i(w)} \right\}.$

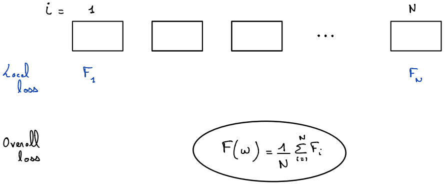


Local loss F_1 F_N

Overall loss $F(w) = \frac{1}{N} \sum_{i=1}^{N} F_i$

General Federated Learning framework

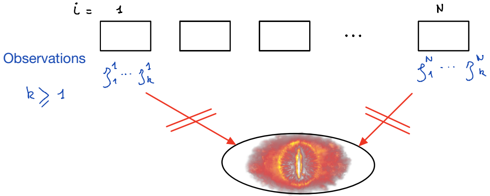
Learning from a set of N agents: $$\min_{w \in \mathbb{R}^d} \left\{ F(w) := \frac{1}{N} \sum_{i=1}^{N} \underbrace{\mathbb{E}_{z \sim \mathscr{D}_i} [\ell(z, w)]}_{F_i(w)} \right\}.$$



→ **4 major challenges.**

General Federated Learning framework

Learning from a set of N agents: $\min_{w \in \mathbb{R}^d} \left\{ F(w) := \frac{1}{N} \sum_{i=1}^{N} \underbrace{\mathbb{E}_{z \sim \mathscr{D}_i} [\ell(z, w)]}_{F_i(w)} \right\}.$



→ **4 major challenges.**

Privacy

Learning from a set of N agents: $\min_{w \in \mathbb{R}^d} \left\{ F(w) := \frac{1}{N} \sum_{i=1}^{N} \underbrace{\mathbb{E}_{z \sim \mathcal{D}_i} \left[\ell(z, w) \right]}_{F_i(w)} \right\}.$

$i = 1$ ... N

Distribution $z_k^1 \sim \mathcal{D}_1$ ... $z_k^N \sim \mathcal{D}_N$

$$\mathcal{D}_i \neq \mathcal{D}_j$$

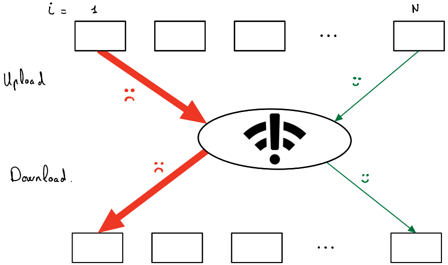
→ **4 major challenges.**

Privacy

Non i.i.d. agents

General Federated Learning framework

Learning from a set of N agents: $\min_{w \in \mathbb{R}^d} \left\{ F(w) := \frac{1}{N} \sum_{i=1}^{N} \underbrace{\mathbb{E}_{z \sim \mathcal{D}_i} \left[\ell(z, w) \right]}_{F_i(w)} \right\}.$



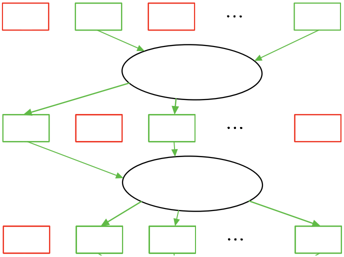
→ **4 major challenges.**

Privacy

Non i.i.d. agents

Optimization with bandwidth constraints

Learning from a set of N agents: $\min_{w \in \mathbb{R}^d} \left\{ F(w) := \frac{1}{N} \sum_{i=1}^{N} \underbrace{\mathbb{E}_{z \sim \mathcal{D}_i} [\ell(z, w)]}_{F_i(w)} \right\}.$

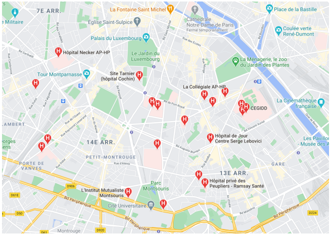


→ **4 major challenges.**

Privacy | **Non i.i.d. agents** | **Optimization with bandwidth constraints** | **Partial participation**

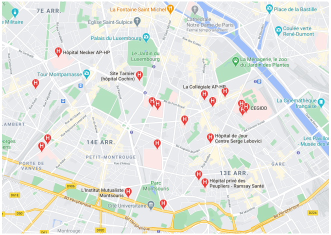
Two Classical Examples

Collaboration between hospitals:



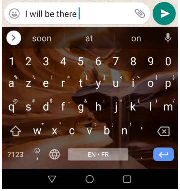
Map of the hospitals in 13-14th arrondissements

Collaboration between hospitals:

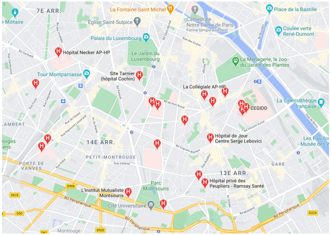


Map of the hospitals in 13-14th arrondissements

Building a collaborative and personalized text model:

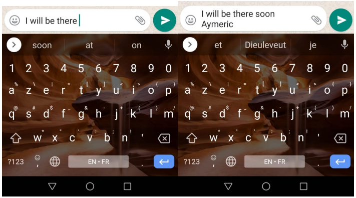


Collaboration between hospitals:



Map of the hospitals in 13-14th arrondissements

Building a collaborative and personalized text model:



`Artemis`: a framework for bi-compression in heterogeneous settings

Our framework

Goal: Learn a **consensus** $w_* = \arg\min F(w)$.

Algorithm: Stochastic Gradient Descent (SGD):

- We iteratively build a sequence of models $(w_k)_{k \geq 0}$.
- Worker i can compute an unbiased estimate g_k^i of the gradient of F_i at the current point w_{k-1}: e.g., $g_k^i := \nabla_w \ell(w_{k-1}, z_k^i)$.
- The central server can update the model computing: $w_k = w_{k-1} - \gamma \frac{1}{N} \sum_{i=1}^N g_k^i$.

Our framework

Goal: Learn a **consensus** $w_* = \arg\min F(w)$.

Algorithm: Stochastic Gradient Descent (SGD):

- We iteratively build a sequence of models $(w_k)_{k \geq 0}$.
- Worker i can compute an unbiased estimate g_k^i of the gradient of F_i at the current point w_{k-1}: e.g., $g_k^i := \nabla_w \ell(w_{k-1}, z_k^i)$.
- The central server can update the model computing: $w_k = w_{k-1} - \gamma \frac{1}{N} \sum_{i=1}^N g_k^i$.

4 challenges / constraints:

0. potentially large group of N agents, with high dimensional data,
1. bandwidth constraints
2. potentially with inactive agents at certain iterations
3. distribution shift between agents
4. "weak" assumptions on the noise on the gradients estimates

In the following, we will enumerate 4 assumptions.

Compression

Several papers considered unidirectional compression, only from the workers to the server.

- Relies on the assumption that the communication cost is higher from the workers to the central node than in the other direction.

Several papers considered unidirectional compression, only from the workers to the server.

- Relies on the assumption that the communication cost is higher from the workers to the central node than in the other direction.

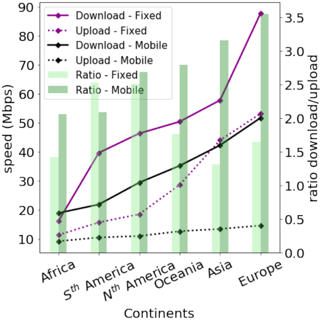


Figure 1: Upload/download speed (in Mbps) for mobile and fixed broadband on left axe. The dataset is gathered from *Speedtest.net*

Bi-directional compression

To limit the number of bits exchanged, we **compress** each signal before transmitting it.

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

Assumption 1

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

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

Several well-known compression operator: quantization, sparsification, top-k coordinates.

↬ Assumption on the compression operator & compression level

Definition 1 (s-quantization operator)

Given $\Delta \in \mathbb{R}^d$, the s-quantization operator $\mathscr{C}_s$ is defined by:

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

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

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

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

Bi-directional compression

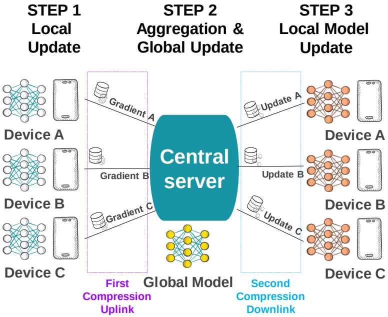


Figure 2: The mechanism of bi-directional compression. First we compress the gradients sent from remote devices, secondly we compress the average of compressed gradient that will be broadcast by the server.

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

Non identically distributed agents

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

Assumption 2

For all $i \in [N]$*:*

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

Non identically distributed agents

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. $h_k^i \to_{k\to\infty} \nabla F_i(w_*)$.

Non identically distributed agents

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:

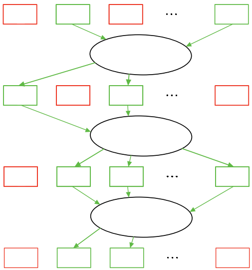
$$w_k = w_{k-1} - \gamma \mathcal{C}_{\mathrm{down}}\left(\frac{1}{N}\sum_{i=1}^{N} \mathcal{C}_{\mathrm{up}}(g_k^i - h_k^i) + h_k^i\right)$$

$$h_{k+1}^i = h_k^i + \alpha \mathcal{C}_{\mathrm{up}}(g_k^i - h_k^i)$$

Device sampling (new!)

Motivation: In practice, some workers may be unavailable / switched off.

w_k model at iteration k.
$\mathscr{C}_{\mathrm{down}}$, $\mathscr{C}_{\mathrm{up}}$ compression operators.
h_k^i memory term and g_k^i gradient.
α learning rate for the memory,
γ step size for the training.



$\Rightarrow$ The update equation becomes:

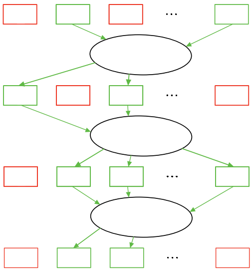
$$w_k = w_{k-1} - \gamma \mathscr{C}_{\mathrm{down}} \left(\frac{1}{pN} \sum_{i \in S_k} \mathscr{C}_{\mathrm{up}}(g_k^i - h_k^i) + h_k^i \right)$$

$$h_{k+1}^i = h_k^i + \alpha \mathscr{C}_{\mathrm{up}}(g_k^i - h_k^i)$$

We maintain the same models on all active workers by broadcasting the updates they have missed.

Motivation: In practice, some workers may be unavailable / switched off.

w_k model at iteration k.
$\mathscr{C}_{\text{down}}$, $\mathscr{C}_{\text{up}}$ compression operators.
h_k^i memory term and g_k^i gradient.
α learning rate for the memory,
γ step size for the training.



$\Rightarrow$ The update equation becomes:

$$w_k = w_{k-1} - \gamma \mathscr{C}_{\text{down}} \left(\frac{1}{pN} \left(\sum_{i \in S_k} \mathscr{C}_{\text{up}}(g_k^i - h_k^i) \right) + h_k \right)$$

$$h_{k+1}^i = h_k^i + \alpha \mathscr{C}_{\text{up}}(g_k^i - h_k^i), \quad h_k = \frac{1}{N} \sum_{i=1}^{N} h_k^i$$

We maintain the same models on all active workers by broadcasting the updates they have missed.

Variance on the noise

Classical assumption: **uniformly bounded variance:**

$$\forall k \geq 1, \forall i \in [N], \quad \mathbb{E}\left[\left\|g_k^i(w_k) - \nabla F_i(w_k)\right\|^2\right] \leq \sigma^2.$$

Assumption 3

Bounded variance at the optimal point:

$$\forall k \geq 1, \forall i \in [N], \quad \mathbb{E}\left[\left\|g_k^i(w_*) - \nabla F_i(w_*)\right\|^2\right] \leq \sigma_*^2.$$

Important in the interpolation regime and because the uniform one is not valid for Least Squares regression !

Related work

Table 1: Relationship with other papers

	QSGD [1]	Diana [4]	Dore [2]	Double Squeeze [6]	Dist EF-SGD [7]	**Artemis (new) [5]**
Data	i.i.d	non i.i.d	i.i.d	i.i.d	i.i.d	non i.i.d
Bounded variance	**Uniformly**	**Uniformly**	**Uniformly**	**Uniformly**	**Uniformly**	**At optimal point**
Compression	One-way	One-way	Two-way	Two-way	Two-way	Two-way
Error compensation			✓	✓	✓	
Memory		✓	✓			✓
Device sampling						✓

Theorem 1 (Convergence of `Artemis`)

For a step size γ, for a learning rate α and for any k in $\mathbb{N}$,

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

with

Variant	E	C
$\alpha = 0$	$(\omega_{\mathcal{C}}^{\text{down}}+1)\left((\omega_{\mathcal{C}}^{\text{up}}+1)\sigma_*^2 + (\omega_{\mathcal{C}}^{\text{up}}+1)B^2\right)$	0
$\alpha \neq 0$	$\sigma_*^2\left((2\omega_{\mathcal{C}}^{\text{up}}+1)(\omega_{\mathcal{C}}^{\text{down}}+1)\right)$	> 0

and $\alpha(\omega_{\mathcal{C}}^{\text{up}}+1) = 1/2$ in the second line

- Linear rate up to a constant of the order of E
- Memory ($\alpha \neq 0$) is needed to obtain linear convergence when $\sigma_*^2 = 0$, in the non i.i.d. case, $B^2 \neq 0$.
- Recovers classical SGD rate in the absence of compression.
- The limit variance increases with the compression level.
- See paper for impact of p

Sketch of the proof

We define a Lyapunov function V_k [as in 4], with k in $[\![1, K]\!]$ and p in $\mathbb{R}^*$:

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

The second part of the Lyapunov corresponds to the memory term: it is the distance between the next element prediction h_k^i and the true gradient $h_*^i = \nabla F_i(w_*)$.

We want to prove that is is a $(1 - \gamma\mu)$ contraction, we need to:

1. Get a first bound on $\| w_k - w_* \|^2$
2. Find a recurrence over the memory term $\left\| h_k^i - h_*^i \right\|^2$
3. Combines the two equations using regularity assumptions:

$$\mathbb{E} V_{k+1} \leq (1 - \gamma\mu) \mathbb{E} V_k + 2\gamma^2 \frac{E}{N}$$

More general convergence

Theorem 2

Sublinear convergence rate for non-strongly convex functions.

Matching lower bound

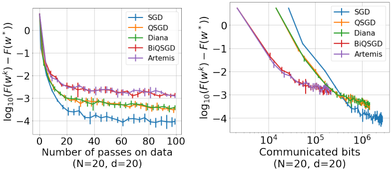
Theorem 3

Lower bound on the asymptotic variance. For a constant step size, the distribution of the iterates converges (in $\mathcal{W}_2$ distance) to a limit distribution which variance matches the upper bound.

Conclusions:

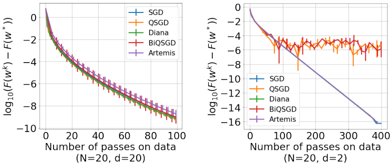
- `Artemis` provides provable reduction of the communication budget for a low precision threshold, and comes with tight guarantees.
- The noise variance at the optimal point is the meaningful quantity.
- For high-precision regimes, Double compression can become less efficient than vanilla SGD.

Experiments : 1 -Numerical validation of the results



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

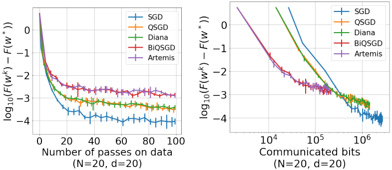
Figure 3: Illustration of `Artemis` compared to existing algorithms on i.i.d. data.



(a) LSR (i.i.d.) **(b)** LR (non-i.i.d.)

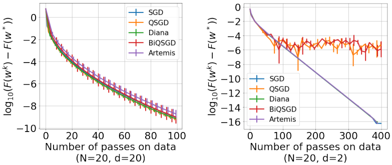
Figure 4: Illustration of the memory benefits when $\sigma_* = 0$: i.i.d. *vs* non-i.i.d.

Experiments : 1 -Numerical validation of the results



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

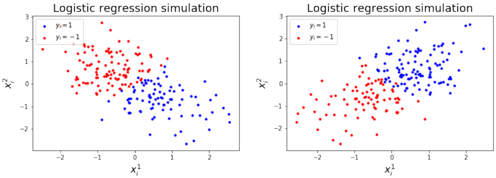
Figure 3: Illustration of `Artemis` compared to existing algorithms on i.i.d. data.



(a) LSR (i.i.d.) **(b)** LR (non-i.i.d.)

Figure 4: Illustration of the memory benefits when $\sigma_* = 0$: i.i.d. *vs* non-i.i.d.

Group heterogeneity:



(a) Distribution 1 **(b)** Distribution 2

Experiments : 2 -Real Datasets

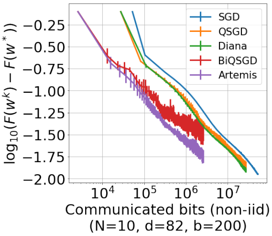


Figure 6: Superconduct (LSR), $b = 200$ $(1000$ iter.$)$

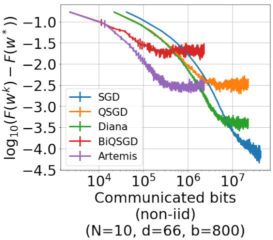


Figure 7: Quantum (LR), $b = 800$ $(1000$ iter.$)$

Experiments : 2 -Real Datasets

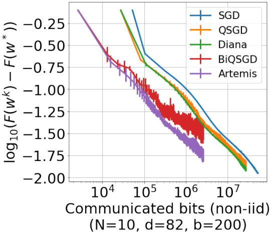


Figure 6: Superconduct (LSR), $b = 200$ (1000 iter.)

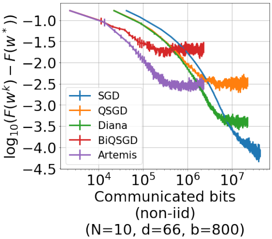


Figure 7: Quantum (LR), $b = 800$ (1000 iter.)

Group heterogeneity:

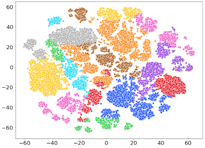


Figure 8: TSNE representation for *quantum*

Reducing the impact of downlink compression: `MCM`

Perturbed iterate point of view

Artemis:

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

MCM: key idea - **preserve the model on the central server.**

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

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

1. Gradient is taken at a random point $\hat{w}_k$ s.t. $\mathbb{E}[\hat{w}_k | w_k] = w_k$
2. Not realistic as it is: `Ghost` algorithm

Convergence for Ghost

1. Control the variance of the local iterate

Theorem 4 (Variance of the local iterates, Ghost)

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

Convergence for Ghost

1. Control the variance of the local iterate

Theorem 4 (Variance of the local iterates, Ghost)

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

2. Deduce convergence of the iterate sequence

Proof technique: Perturbed iterate analysis [3]

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

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

1. Control the variance of the local iterate

Theorem 4 (Variance of the local iterates, `Ghost`)

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

2. Deduce convergence of the iterate sequence

Theorem 5 (Contraction for `Ghost`, convex case)

For smooth & convex objective, bounded variance (uniform), if $\gamma L(1+\omega_{\mathcal{C}}^{\text{up}}/N) \leq \frac{1}{2}$.

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

Corollary 6 (Convergence of `Ghost`, convex case)

For a given step size $\gamma = 1/(L\sqrt{K})$*, after running* K *in* $\mathbb{N}$ *iterations, we have, for* $\bar{w}_K = K^{-1}\sum_{i=1}^{K} w_i$*:*

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

with $\Phi = (1 + \omega_{\mathcal{C}}^{\mathrm{up}}) \left(1 + 2\frac{\omega_{\mathcal{C}}^{\mathrm{down}}}{\sqrt{K}}\right)$.

Implementable algorithms ?

Simplest solution:

$$\begin{cases} w_{k+1} = w_k - \gamma \frac{1}{N} \sum_{i=1}^N \mathscr{C}_{\mathrm{up}} \left(g_{k+1}^i (\hat{w}_k) \right) . \\ \hat{w}_{k+1} = C_{\mathrm{down}} (w_{k+1}) \end{cases}$$

Implementable algorithms ?

Simplest solution:

$$\begin{cases} w_{k+1} = w_k - \gamma \frac{1}{N} \sum_{i=1}^{N} \mathscr{C}_{\mathrm{up}} \left(g_{k+1}^i (\hat{w}_k) \right) . \\ \hat{w}_{k+1} = C_{\mathrm{down}}(w_{k+1}) \end{cases}$$

Compress difference $w_{k+1} - \widehat{w}_k$

$$\begin{cases} w_{k+1} = w_k - \gamma \frac{1}{N} \sum_{i=1}^{N} \mathscr{C}_{\mathrm{up}} \left(g_{k+1}^i (\hat{w}_k) \right) . \\ \hat{w}_{k+1} = \hat{w}_k + C_{\mathrm{down}}(w_{k+1} - \hat{w}_k) \end{cases}$$

Simplest solution:

$$\begin{cases} w_{k+1} = w_k - \gamma \frac{1}{N} \sum_{i=1}^N \mathscr{C}_{\mathrm{up}} \left(g_{k+1}^i(\hat{w}_k) \right) . \\ \hat{w}_{k+1} = C_{\mathrm{down}}(w_{k+1}) \end{cases}$$

Compress difference $w_{k+1} - \widehat{w}_k$

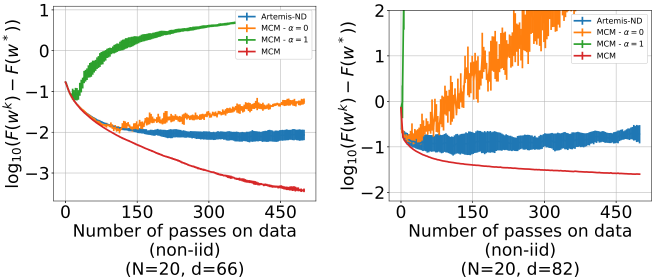
$$\begin{cases} w_{k+1} = w_k - \gamma \frac{1}{N} \sum_{i=1}^N \mathscr{C}_{\mathrm{up}} \left(g_{k+1}^i(\hat{w}_k) \right) . \\ \hat{w}_{k+1} = \hat{w}_k + C_{\mathrm{down}}(w_{k+1} - \hat{w}_k) \end{cases}$$

$\looparrowright$ **add a downlink memory term** $(H_k)_k$,

$$\begin{cases} \Omega_{k+1} = w_{k+1} - H_k \,, \\ \widehat{w}_{k+1} = H_k + \mathscr{C}_{\mathrm{down}}(\Omega_{k+1}) \\ H_{k+1} = H_k + \alpha \mathscr{C}_{\mathrm{down}}(\Omega_{k+1}). \end{cases}$$

2. Deduce convergence of the iterate sequence

Comparison between the three variants



(a) Quantum - $b = 400$. **(b)** Superconduct - $b = 50$.

Figure 9: Comparing `MCM` with three other algorithms using a non-degraded update, $\gamma = 1/L$. `Artemis-ND` stands for `Artemis` with a non-degraded update. Best seen in colors.

Convergence rate `MCM` algorithm

1. Control the variance of the local iterate

Theorem 7

Consider the `MCM` update. If $\gamma \leq 1/(8\omega_{\mathcal{C}}^{\mathrm{down}} L)$ and $\alpha \leq 1/(4\omega_{\mathcal{C}}^{\mathrm{down}})$, for $k \in \mathbb{N}$:

$$\mathbb{E}[\| w_k - \widehat{w}_k \|^2] \leq \gamma^2 \omega_{\mathcal{C}}^{\mathrm{down}} \left(\frac{4\sigma^2 (1 + \omega_{\mathcal{C}}^{\mathrm{up}})}{Nb\alpha} + 2 \left(\frac{1}{\alpha} + \frac{\omega_{\mathcal{C}}^{\mathrm{up}}}{N} \right) \sum_{t=1}^{k} \left(1 - \frac{\alpha}{2} \right)^{k-t} \mathbb{E} \| \nabla F(\widehat{w}_{t-1}) \|^2 \right).$$

Convergence rate MCM algorithm

1. Control the variance of the local iterate

Theorem 7

Consider the MCM update. If $\gamma \leq 1/(8\omega_{\mathcal{C}}^{\mathrm{down}} L)$ *and* $\alpha \leq 1/(4\omega_{\mathcal{C}}^{\mathrm{down}})$, *for* $k \in \mathbb{N}$*:*

$$\mathbb{E}[\| w_k - \widehat{w}_k \|^2] \leq \gamma^2 \omega_{\mathcal{C}}^{\mathrm{down}} \left(\frac{4\sigma^2(1+\omega_{\mathcal{C}}^{\mathrm{up}})}{Nb\alpha} + 2\left(\frac{1}{\alpha} + \frac{\omega_{\mathcal{C}}^{\mathrm{up}}}{N}\right) \sum_{t=1}^{k} \left(1 - \frac{\alpha}{2}\right)^{k-t} \mathbb{E}\|\nabla F(\widehat{w}_{t-1})\|^2 \right).$$

2. Deduce convergence of the iterate sequence

Theorem 8 (Convergence of MCM)

For a given K *in* $\mathbb{N}$ *large enough, a step size* $\gamma = 1/(L\sqrt{K})$, *a given learning rate* $\alpha = 1/(8\omega_{\mathcal{C}}^{\mathrm{down}})$, *after running* K *iterations, we have:*

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

with $\Phi = (1+\omega_{\mathcal{C}}^{\mathrm{up}})\left(1 + \frac{64(\omega_{\mathcal{C}}^{\mathrm{down}})^2}{\sqrt{K}}\right)$.

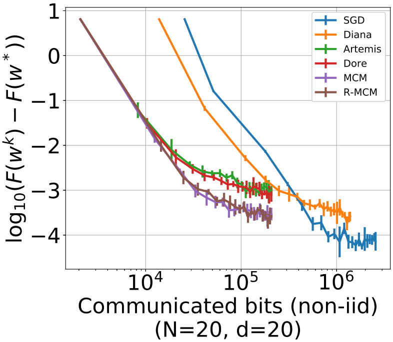
Conclusion

Extensions:

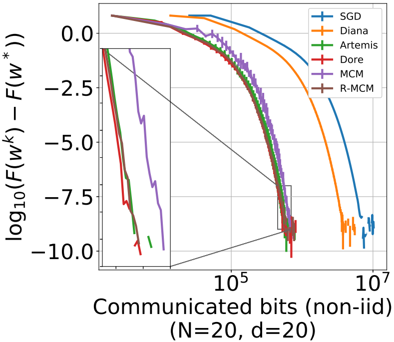
1. Convergence in the strongly-convex, non convex cases.
2. Worker dependent compression: `Rand-MCM`

$$\widehat{w}_{k+1}^{i} = H_k^{i} + \mathscr{C}_{\mathrm{down}}^{i}(w_{k+1} - H_k^{i})$$

- Useful with partial participation
- Memory limitation
- Improves the convergence rate (on quadratics)
- Business applications

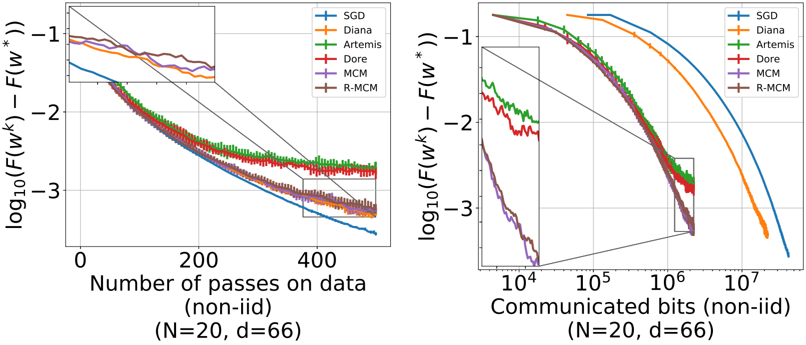


(a) LSR: $\sigma^2 \neq 0,\ \gamma = (L\sqrt{k})^{-1}$



(b) LSR: $\sigma^2 = 0,\ \gamma = 1/L$

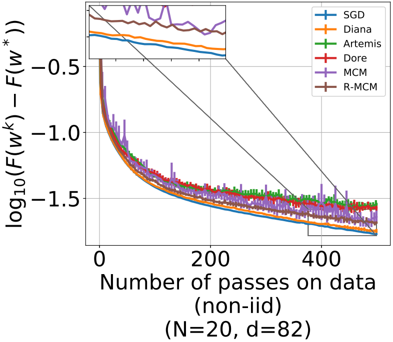
Figure 10: Toy dataset, X axis in # bits.



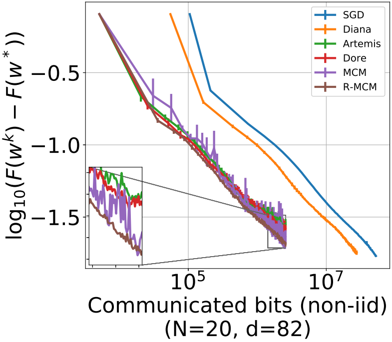
(a) X axis in # iterations

(b) X axis in # bits

Figure 11: Quantum with $b = 400$, $\gamma = 1/L$ (LSR).



(a) X axis in # iterations



(b) X axis in # bits

Figure 12: Superconduct with $b = 50$, $\gamma = 1/L$ (LR).

Conclusion

Take home message

1. New algorithm for bi-directional compression:
    - *preserved* central model.
    - relying on memory trick on the downlink communication
2. Reduces (nearly cancels) impact of downlink compression
3. Achieves the same rate of convergence as unidirectional compression.
4. `Rand-MCM` framework enables multiple possible extensions.

Open questions

1. Even faster ? no dependence in ω_{down}?
2. Variance reduced modification.
3. Proofs with partial participation.

Thank you for your attention :)

Bi-directional compression for Federated Learning: `Artemis` & `MCM`

Aymeric Dieuleveut
CMAP, École Polytechnique, Institut Polytechnique de Paris

Joint work with Constantin Philippenko

References:

- Artemis paper
- MCM paper

References

[1] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding. *arXiv:1610.02132 [cs]*, December 2017. arXiv: 1610.02132.

[2] Xiaorui Liu, Yao Li, Jiliang Tang, and Ming Yan. A Double Residual Compression Algorithm for Efficient Distributed Learning. *arXiv:1910.07561 [cs, stat]*, October 2019. arXiv: 1910.07561.

[3] Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. 07 2015.

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

[5] Constantin Philippenko and Aymeric Dieuleveut. Artemis: tight convergence guarantees for bidirectional compression in federated learning.

[6] Hanlin Tang, Xiangru Lian, Chen Yu, Tong Zhang, and Ji Liu. DoubleSqueeze: Parallel Stochastic Gradient Descent with Double-Pass Error-Compensated Compression. *arXiv:1905.05957 [cs]*, June 2019. arXiv: 1905.05957.

[7] 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 AlchÃľ-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 11450–11460. Curran Associates, Inc., 2019.