Bi-directional compression for Federated Learning: Artemis & MCM

Aymeric Dieuleveut
CMAP, École Polytechnique, Institut Polytechnique de Paris

Joint work with Constantin Philippenko
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: $\min_{w \in \mathbb{R}^d} \left\{ F(w) := \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{z \sim D_i} [\ell(z, w)] \right\}$. 

1. Privacy
2. Non i.i.d. agents
3. Optimization with bandwidth constraints
4. Partial participation
General Federated Learning framework

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

\[ F(w) = \frac{1}{N} \sum_{i=1}^{N} F_i(w) \]

- **Privacy**
- **Non i.i.d. agents**
- **Optimization with bandwidth constraints**
- **Partial participation**
General Federated Learning framework

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

→ 4 major challenges.
General Federated Learning framework

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

→ 4 major challenges.

Privacy
General Federated Learning framework

Learning from a set of $N$ agents: 

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

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

$\rightarrow$ 4 major challenges.

Privacy

Non i.i.d. agents
General Federated Learning framework

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

→ 4 major challenges.

Privacy  Non i.i.d. agents  Optimization with bandwidth constraints
General Federated Learning framework

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

→ 4 major challenges.

<table>
<thead>
<tr>
<th>Privacy</th>
<th>Non i.i.d. agents</th>
<th>Optimization with bandwidth constraints</th>
<th>Partial participation</th>
</tr>
</thead>
</table>
Two Classical Examples

Collaboration between hospitals:

Map of the hospitals in 13-14th arrondissements
Two Classical Examples

Collaboration between hospitals:

Map of the hospitals in 13-14th arrondissements

Building a collaborative and personalized text model:
Two Classical Examples

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** $w_\ast = \arg\min F(w)$.

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

4 challenges / constraints:
- potentially large group of $N$ agents, with high dimensional data,
- bandwidth constraints
- potentially with inactive agents at certain iterations
- distribution shift between agents
- “weak” assumptions on the noise on the gradients estimates

In the following, we will enumerate 4 assumptions.
Our framework

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

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

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

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

**Assumption 1**

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

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

Several well-known compression operators: 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 $C_s$ is defined by:

$$C_s(\Delta) := \text{sign}(\Delta) \times \|\Delta\|_2 \times \frac{\psi}{s}.$$  

$\psi \in \mathbb{R}^d$ 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.

⇒ The update equation becomes: \[ w_k = w_{k-1} - \gamma c_{\text{down}} \left( \frac{1}{N} \sum_{i=1}^{N} c_{\text{up}}(g^i_k) \right) \]
Non identically distributed agents

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

**Assumption 2**

For all $i \in [N]$:

$$\|\nabla F_i(w_*)\|^2 \leq B^2$$
**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" $h^i_k$ s.t. $h^i_k \rightarrow_{k \rightarrow \infty} \nabla F_i(w_*)$. 

$$w_k = w_{k-1} - \gamma C_{\text{down}} \left( \frac{1}{N} \sum_{i=1}^{N} C_{\text{up}} (g^i_k - h^i_k) + h^i_k \right)$$
**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” $h^i_k$ s.t. $h^i_k \to_{k \to \infty} \nabla F_i(w_*)$.

$\Rightarrow$ The update equation becomes:

\[
 w_k = w_{k-1} - \gamma \mathcal{C}_{\text{down}} \left( \frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\text{up}} (g^i_k - h^i_k) + h^i_k \right)
\]

\[
 h^i_{k+1} = h^i_k + \alpha \mathcal{C}_{\text{up}} (g^i_k - h^i_k)
\]
**Motivation:** In practice, some workers may be unavailable / switched off.

$w_k$ model at iteration $k$.

$C_{\text{down}}$, $C_{\text{up}}$ compression operators.

$h_k^i$ memory term and $g_k^i$ gradient.

$\alpha$ learning rate for the memory,

$\gamma$ step size for the training.

⇒ The update equation becomes:

\[
\begin{align*}
    w_k &= w_{k-1} - \gamma C_{\text{down}} \left( \frac{1}{pN} \sum_{i \in S_k} C_{\text{up}} (g_k^i - h_k^i) + h_k^i \right) \\
    h_{k+1}^i &= h_k^i + \alpha C_{\text{up}} (g_k^i - h_k^i)
\end{align*}
\]

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 \( k \).
\( C_{\text{down}}, C_{\text{up}} \) compression operators.
\( h_k^i \) memory term and \( g_k^i \) gradient.
\( \alpha \) learning rate for the memory,
\( \gamma \) step size for the training.

\[ \Rightarrow \text{The update equation becomes:} \]

\[ w_k = w_{k-1} - \gamma C_{\text{down}} \left( \frac{1}{pN} \sum_{i \in S_k} C_{\text{up}} (g_k^i - h_k^i) \right) + h_k \]

\[ h_{k+1}^i = h_k^i + \alpha C_{\text{up}} (g_k^i - h_k^i), \quad h_k = \frac{1}{N} \sum_{i=1}^{N} h_k^i \]

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:**

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

**Assumption 3**

*Bounded variance at the optimal point:*

\[
\forall k \geq 1, \forall i \in [N], \quad \mathbb{E} \left[ \left\| g^i_k(w^*_k) - \nabla F_i(w^*_k) \right\|^2 \right] \leq \sigma^*_2.
\]

Important in the interpolation regime and because the uniform one is not valid for Least Squares regression!
## Table 1: Relationship with other papers

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>i.i.d</td>
<td>non i.i.d</td>
<td>i.i.d</td>
<td>i.i.d</td>
<td>i.i.d</td>
<td>non i.i.d</td>
</tr>
<tr>
<td>Bounded variance</td>
<td>Uniformly</td>
<td>Uniformly</td>
<td>Uniformly</td>
<td>Uniformly</td>
<td>Uniformly</td>
<td>At optimal point</td>
</tr>
<tr>
<td>Compression</td>
<td>One-way</td>
<td>One-way</td>
<td>Two-way</td>
<td>Two-way</td>
<td>Two-way</td>
<td>Two-way</td>
</tr>
<tr>
<td>Error compensation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Memory</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Device sampling</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Theorem 1 (Convergence of Artemis)

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

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

with

<table>
<thead>
<tr>
<th>Variant</th>
<th>$E$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha = 0$</td>
<td>$(\omega_{\text{down}}^\alpha + 1) ((\omega_{\text{up}}^\alpha + 1)\sigma_\star^2 + (\omega_{\text{up}}^\alpha + 1) B^2)$</td>
<td>0</td>
</tr>
<tr>
<td>$\alpha \neq 0$</td>
<td>$\sigma_\star^2 (2\omega_{\text{up}}^\alpha + 1)(\omega_{\text{down}}^\alpha + 1)$</td>
<td>$&gt; 0$</td>
</tr>
</tbody>
</table>

and $\alpha(\omega_{\text{up}}^\alpha + 1) = 1/2$ 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 $\sigma_\star^2 = 0$, 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 $\mathbb{R}^*$:

$$V_k = \|w_k - w_*\|^2 + 2\gamma^2 C \frac{1}{N} \sum_{i=1}^{N} \|h^i_k - h^i_*\|^2.$$

The second part of the Lyapunov corresponds to the memory term: it is the distance between the next element prediction $h^i_k$ 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 $\|h^i_k - h^i_*\|^2$
3. Combines the two equations using regularity assumptions:

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

## 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 \( W_2 \) 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: $\sigma^2 \neq 0$  \hspace{1cm} (b) X-axis in bits

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

(a) LSR (i.i.d.) \hspace{1cm} (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 $\sigma_\ast = 0$: 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:

\[ w_k = w_{k-1} - \gamma \mathcal{C}_{\text{down}} \left( \frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\text{up}}(g^i_k(w_{k-1})) \right) \]

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

\[ w_k = w_{k-1} - \gamma \left( \frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\text{up}}(g^i_k(\hat{w}_{k-1})) \right) \]

\[ \hat{w}_k = w_{k-1} - \gamma \mathcal{C}_{\text{down}} \left( \frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\text{up}}(g^i_k(\hat{w}_{k-1})) \right) \]

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
1. Control the variance of the local iterate

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

\[
\mathbb{E} \left[ \| w_{k-1} - \hat{w}_{k-1} \|^2 \mid \hat{w}_{k-2} \right] \leq \gamma^2 \omega_{\mathcal{C}} \downarrow \left( \frac{(1 + \omega_{\mathcal{C}})^2}{Nb} + \left( 1 + \frac{\omega_{\mathcal{C}}}{N} \right) \| \nabla F(\hat{w}_{k-2}) \|^2 \right).
\]
Convergence for Ghost

1. Control the variance of the local iterate

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

\[
\mathbb{E} \left[ \| w_{k-1} - \hat{w}_{k-1} \|^2 \mid \hat{w}_{k-2} \right] \leq \gamma^2 \omega_c^{\text{down}} \left( \frac{(1 + \omega_c^{\text{up}})\sigma^2}{Nb} + \left( 1 + \frac{\omega_c^{\text{up}}}{N} \right) \| \nabla F(\hat{w}_{k-2}) \|^2 \right).
\]

2. Deduce convergence of the iterate sequence

Proof technique: Perturbed iterate analysis [3]

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

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

1. Control the variance of the local iterate

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

\[
\mathbb{E} [\| w_{k-1} - \hat{w}_{k-1} \|^2 \mid \hat{w}_{k-2}] \leq \gamma^2 \omega^\down_\mathcal{C} \left( \frac{(1 + \omega^\up_\mathcal{C}) \sigma^2}{Nb} + \left(1 + \frac{\omega^\up_\mathcal{C}}{N}\right) \| \nabla F(\hat{w}_{k-2}) \|^2 \right).
\]

2. Deduce convergence of the iterate sequence

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

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

\[
\mathbb{E} \| w_k - w_* \|^2 \leq \mathbb{E} \| w_{k-1} - w_* \|^2 - \gamma \mathbb{E} (F(w_{k-1}) - F_*) - \frac{\gamma}{2L} \mathbb{E} \| \nabla F(\hat{w}_{k-1}) \|^2
\]

\[
+ 2\gamma^3 \omega^\down_\mathcal{C} L \left(1 + \frac{\omega^\up_\mathcal{C}}{N}\right) \mathbb{E} \| \nabla F(\hat{w}_{k-2}) \|^2 + \gamma^2 \frac{(1 + \omega^\up_\mathcal{C}) \sigma^2}{Nb} \left(1 + 2\gamma L \omega^\down_\mathcal{C}\right).
\]
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 $\Phi = (1 + \omega_{\mathcal{E}}^{\text{up}}) \left(1 + 2 \frac{\omega_{\mathcal{E}}^{\text{down}}}{\sqrt{K}}\right)$. 
Implementable algorithms?

Simplest solution:

\[
\begin{align*}
  w_{k+1} &= w_k - \gamma \frac{1}{N} \sum_{i=1}^{N} C_{\text{up}}(g^i_{k+1}(\hat{w}_k)) \\
  \hat{w}_{k+1} &= C_{\text{down}}(w_{k+1})
\end{align*}
\]
Implementable algorithms?

Simplest solution:

\[
\begin{align*}
    w_{k+1} &= w_k - \gamma \frac{1}{N} \sum_{i=1}^{N} C_{\text{up}}\left(g^i_{k+1}(\hat{w}_k)\right). \\
    \hat{w}_{k+1} &= C_{\text{down}}(w_{k+1})
\end{align*}
\]

Compress difference \( w_{k+1} - \hat{w}_k \)

\[
\begin{align*}
    w_{k+1} &= w_k - \gamma \frac{1}{N} \sum_{i=1}^{N} C_{\text{up}}\left(g^i_{k+1}(\hat{w}_k)\right). \\
    \hat{w}_{k+1} &= \hat{w}_k + C_{\text{down}}(w_{k+1} - \hat{w}_k)
\end{align*}
\]
Implementable algorithms?

Simplest solution:

\[
\begin{align*}
\{ \\
  w_{k+1} &= w_k - \gamma \frac{1}{N} \sum_{i=1}^{N} C_{up}(g_{k+1}^i(\hat{w}_k)) \\
  \hat{w}_{k+1} &= C_{down}(w_{k+1})
\end{align*}
\]

Compress difference \( w_{k+1} - \hat{w}_k \)

\[
\begin{align*}
\{ \\
  w_{k+1} &= w_k - \gamma \frac{1}{N} \sum_{i=1}^{N} C_{up}(g_{k+1}^i(\hat{w}_k)) \\
  \hat{w}_{k+1} &= \hat{w}_k + C_{down}(w_{k+1} - \hat{w}_k)
\end{align*}
\]

\( \xrightarrow{} \) add a downlink memory term \((H_k)_k\),

\[
\begin{align*}
\{ \\
  \Omega_{k+1} &= w_{k+1} - H_k, \\
  \hat{w}_{k+1} &= H_k + C_{down}(\Omega_{k+1}) \\
  H_{k+1} &= H_k + \alpha C_{down}(\Omega_{k+1}).
\end{align*}
\]

2. Deduce convergence of the iterate sequence
Comparison between the three variants

\[
\log_{10}(F(w^k) - F(w^*))
\]

(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.
1. Control the variance of the local iterate

Theorem 7

Consider the MCM update. If \( \gamma \leq 1/(8 \omega_e \text{down} L) \) and \( \alpha \leq 1/(4 \omega_e \text{down}) \), for \( k \in \mathbb{N} \):

\[
\mathbb{E}[\| w_k - \hat{w}_k \|^2] \leq \gamma^2 \omega_e \text{down} \left( \frac{4 \sigma^2 (1 + \omega_e \text{up})}{Nb\alpha} \right) + 2 \left( \frac{1}{\alpha} + \frac{\omega_e \text{up}}{N} \right) \sum_{t=1}^{k} \left( 1 - \frac{\alpha}{2} \right)^{k-t} \mathbb{E} \| \nabla F(\hat{w}_{t-1}) \|^2.
\]
Convergence rate MCM algorithm

1. Control the variance of the local iterate

Theorem 7

Consider the MCM update. If $\gamma \leq 1/(8\omega_{\text{up}}L)$ and $\alpha \leq 1/(4\omega_{\text{up}})$, for $k \in \mathbb{N}$:

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

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 $\alpha = 1/(8\omega_{\text{up}})$, after running $K$ iterations, we have:

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

with $\Phi = (1 + \omega_{\text{up}}) \left( 1 + \frac{64(\omega_{\text{up}})^2}{\sqrt{K}} \right)$. 
Conclusion

Extensions:

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

\[ \hat{w}_{k+1}^i = H_k^i + \mathcal{C}_{\text{down}}^i (w_{k+1}^i - H_k^i) \]

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

(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.
Experiments ii

Figure 11: Quantum with $b = 400$, $\gamma = 1/L$ (LSR).
Figure 12: Superconduct with $b = 50, \gamma = 1/L$ (LR).
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.

Open questions

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


