Bidirectional compression for federated learning in heterogeneous setting

Constantin Philippenko
CMAP, École Polytechnique, Institut Polytechnique de Paris
Accenture Labs, Sophia Antipolis

Reviewers:
Mikael Johansson
Jérôme Malick

Examiners:
Manon Costa
Robert Gower
Martin Jaggi
Kevin Scaman

Ph.D. supervisors: Aymeric Dieuleveut and Eric Moulines

Ph.D. defense, September 18th, 2023

Contents

General introduction

Figure 1: Automatic plant identification from photos using the mobile app [Pl@ntNet].

From a simple example to the challenges of my thesis

Figure 1: Automatic plant identification from photos using the mobile app [Pl@ntNet].

Goal of machine learning:

Find a mathematical relationship between the input (here the images) and the output (here the name of the plant).

Figure 1: Automatic plant identification from photos using the mobile app [Pl@ntNet].

Goal of machine learning:

Find a mathematical relationship between the input (here the images) and the output (here the name of the plant).

Paradigm of my thesis: data is not centralized on a single location.

From a simple example to the challenges of my thesis

Figure 1: Automatic plant identification from photos using the mobile app [Pl@ntNet].

Goal of machine learning:

Find a mathematical relationship between the input (here the images) and the output (here the name of the plant).

Paradigm of my thesis: data is not centralized on a single location.

Privacy

Communication cost

Data heterogeneity

From a simple example to the challenges of my thesis

Figure 1: Automatic plant identification from photos using the mobile app [Pl@ntNet].

Goal of machine learning:

Find a mathematical relationship between the input (here the images) and the output (here the name of the plant).

Paradigm of my thesis: data is not centralized on a single location.

Privacy

Communication cost

Data heterogeneity

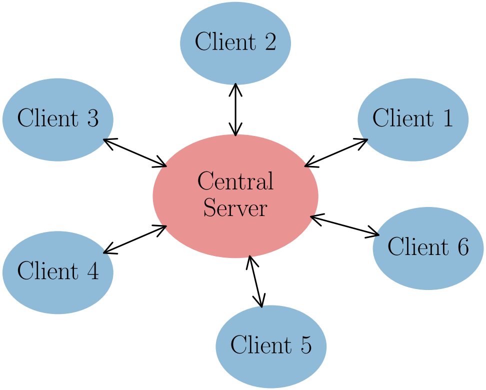
Goal of my thesis:

Focus simultaneously on two challenges: **reducing the cost of communication** *and* considering a **heterogeneous setting**.

Federated learning: an optimization problem

Setting of federated learning:

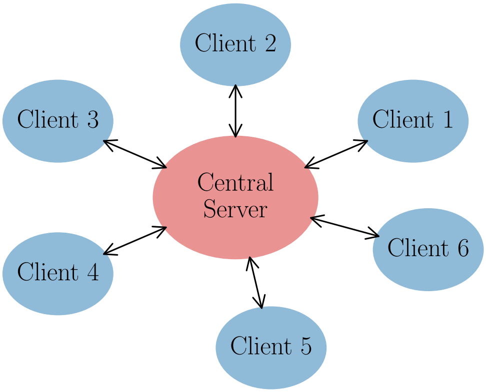
A central server orchestrate the training.



Federated learning: an optimization problem

Setting of federated learning:

A central server orchestrate the training.

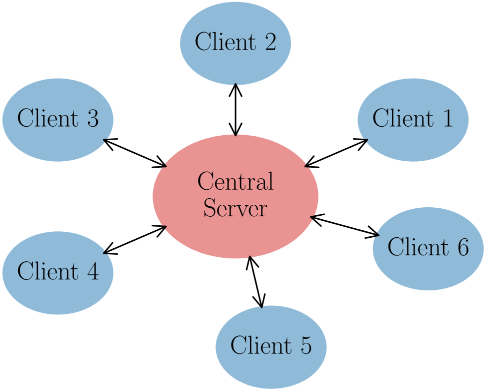


Each client $i \in \mathbb{N}^*$ have access to a “objective function” F_i measuring the error of prediction for a model $w \in \mathbb{R}^d$.

Federated learning: an optimization problem

Setting of federated learning:

A central server orchestrate the training.



Each client $i \in \mathbb{N}^*$ have access to a "objective function" F_i measuring the error of prediction for a model $w \in \mathbb{R}^d$.

We need to find the optimal model w_* such that:

$$w_* = \arg\min_{w \in \mathbb{R}^d} \frac{1}{N} \sum_{i=1}^{N} F_i(w).$$

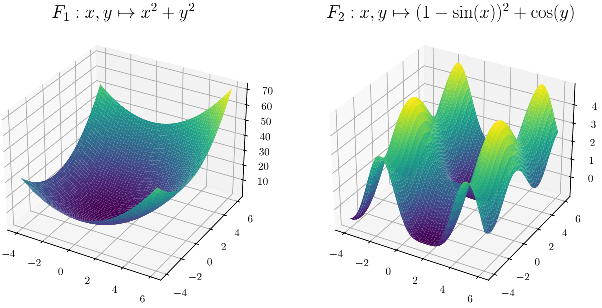
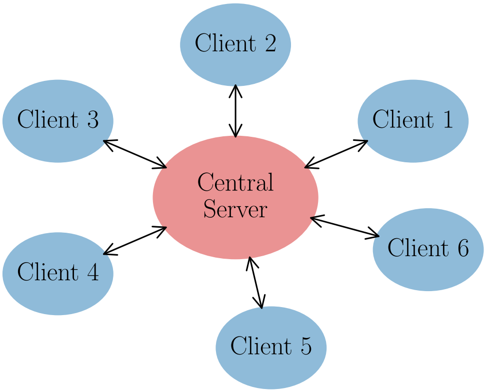


Figure 2: Examples of two objective functions

Setting of federated learning:

A central server orchestrate the training.



Each client $i \in \mathbb{N}^*$ have access to a "objective function" F_i measuring the error of prediction for a model $w \in \mathbb{R}^d$.

We need to find the optimal model w_* such that:

$$w_* = \arg\min_{w \in \mathbb{R}^d} \frac{1}{N} \sum_{i=1}^{N} F_i(w).$$

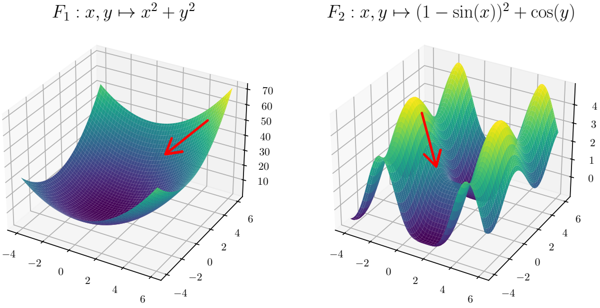


Figure 2: Examples of two objective functions

To find the optimal model w_*, we follow the slope (gradient descent).

Framework for bidirectional compression

Two challenges of Federated Learning

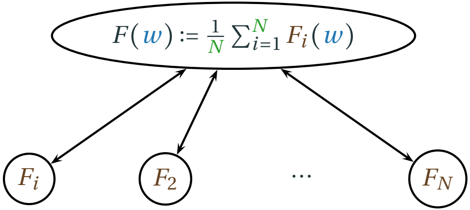
Goal : learning from a set of N clients [MMR+17]

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

F: global cost function
F_i: local loss
N: clients
d: dimension
w: model
$\mathcal{D}_i$: local data distribution

Global loss

Local loss



Distributed SGD: $\forall k \in \mathbb{N}, w_k = w_{k-1} - \gamma\left(\frac{1}{N} \sum_{i=1}^{N} g_k^i(w_{k-1})\right)$.

Two challenges of Federated Learning

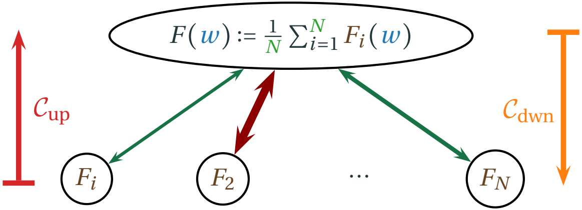
Goal : learning from a set of N clients [MMR+17]

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

F: global cost function
F_i: local loss
N: clients
d: dimension
w: model
$\mathcal{D}_i$: local data distribution

Global loss

Local loss



$\mapsto$ **Challenge 1: reduce communication costs**

Distributed SGD: $\forall k \in \mathbb{N}, w_k = w_{k-1} - \gamma \left(\frac{1}{N} \sum_{i=1}^{N} g_k^i(w_{k-1}) \right)$.

Two challenges of Federated Learning

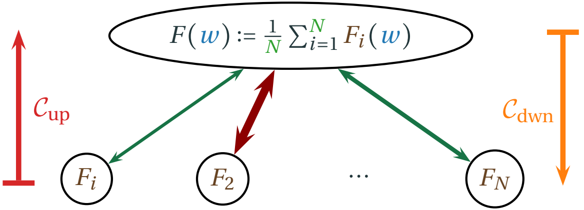
Goal : learning from a set of N clients [MMR+17]

F: global cost function
F_i: local loss
N: clients
d: dimension
w: model
$\mathcal{D}_i$: local data distribution

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

Global loss

Local loss



↦ **Challenge 1: reduce communication costs**

↦ **Challenge 2 handle heterogeneous clients**

Distributed SGD: $\forall k \in \mathbb{N}, w_k = w_{k-1} - \gamma \left(\frac{1}{N} \sum_{i=1}^{N} g_k^i(w_{k-1}) \right).$

Bidirectional compression

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

↦ **Focus on bidirectional compression** [LLTY20, PD20, TYL+19, ZHK19, PD21].

Bidirectional compression

- To limit the number of bits exchanged, we **compress** each signal before transmitting it.
- ↦ **Focus on bidirectional compression** [LLTY20, PD20, TYL+19, ZHK19, PD21].
- ↦ We introduce two compression operators $\mathcal{C}_{\mathrm{dwn}} \downarrow$ and $\mathcal{C}_{\mathrm{up}} \uparrow$.

Compressed distributed SGD:

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

Bidirectional compression

⤼ To limit the number of bits exchanged, we **compress** each signal before transmitting it.
↦ **Focus on bidirectional compression** [LLTY20, PD20, TYL+19, ZHK19, PD21].
↦ We introduce two compression operators $\mathcal{C}_{\mathrm{dwn}} \downarrow$ and $\mathcal{C}_{\mathrm{up}} \uparrow$.
Compressed distributed SGD:

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

Assumption 1 (One assumption to rule them all)

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* z *in* $\mathbb{R}^d$ *:*

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

The compressors are said to be *Unbiased with a Relatively Bounded Variance* (URBV).

Exemples of compressors

1. Sparsification based:
   - Rand-k: keeps k coordinates,
   - p-Sparsification: keeps each coordinate with probability p,
   - p-partial participation: sends the complete vector with probability p,
   - Sketching: using a random projection matrix into a lower-dimension space.
2. Quantization based on a codebook:
   - (Stabilized) scalar quantization (coordinate compressed independently),
   - Delaunay quantization.

Impact of heterogeneity

Compressed distributed SGD: $\forall k \in \mathbb{N}, w_k = w_{k-1} - \gamma \mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\mathrm{up}} \left(g_k^i (w_{k-1}) \right) \right)$.

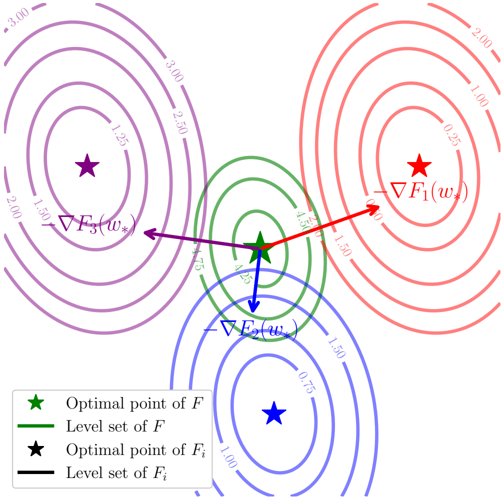
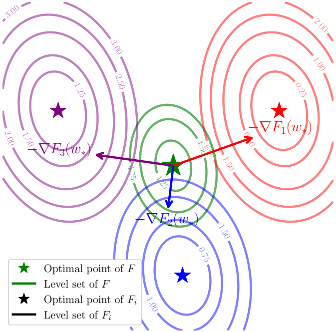


Figure 3: Illustration of heterogeneity on three clients, the objective functions are quadratic. We represent the optimal points, the level set, and the opposite gradient at the optimal point.

From a first theorem to a glance at contributions

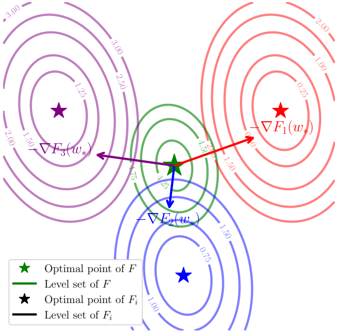
Compressed distributed SGD: $\forall k \in \mathbb{N}, w_k = w_{k-1} - \gamma \mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^N \mathcal{C}_{\mathrm{up}} (g_k^i (w_{k-1})) \right)$.



Assumption 2 (Bounded gradient at w_*)

There exists an optimal parameter w_* minimizing F (not necessarily unique) and a constant $B \in \mathbb{R}_+$, such that $\frac{1}{N} \sum_{i=1}^N \| \nabla F_i(w_*) \|^2 = B^2$.

Compressed distributed SGD: $\forall k \in \mathbb{N}, w_k = w_{k-1} - \gamma \mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^N \mathcal{C}_{\mathrm{up}} (g_k^i (w_{k-1})) \right)$.



Assumption 2 (Bounded gradient at w_*)

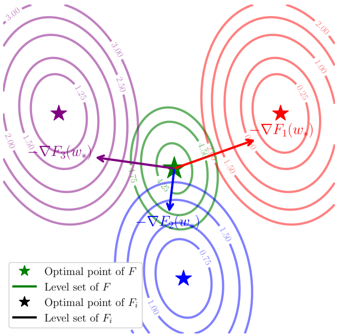
There exists an optimal parameter w_* minimizing F (not necessarily unique) and a constant $B \in \mathbb{R}_+$, such that $\frac{1}{N} \sum_{i=1}^N \| \nabla F_i (w_*) \|^2 = B^2$.

Assumption 3 (Noise over stochastic gradients computation)

The noise over stochastic gradients is zero-centered and its variance is uniformly bounded by a constant $\sigma \in \mathbb{R}_+$, such that for all k in $\mathbb{N}$, for all z in $\mathbb{R}^d$ we have: $\mathbb{E}[\| g_k(z) - \nabla F(z) \|^2] \leq \sigma^2$.

From a first theorem to a glance at contributions

Compressed distributed SGD: $\forall k \in \mathbb{N}, w_k = w_{k-1} - \gamma \mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^N \mathcal{C}_{\mathrm{up}} \left(g_k^i (w_{k-1}) \right) \right).$



Assumption 2 (Bounded gradient at w_*)

There exists an optimal parameter w_* minimizing F (not necessarily unique) and a constant $B \in \mathbb{R}_+$, such that $\frac{1}{N} \sum_{i=1}^N \|\nabla F_i(w_*)\|^2 = B^2$.

Assumption 3 (Noise over stochastic gradients computation)

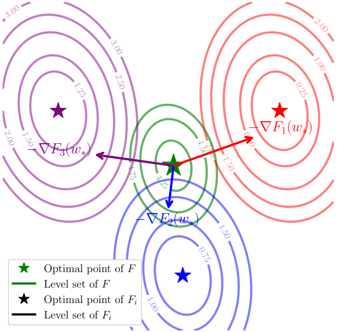
The noise over stochastic gradients is zero-centered and its variance is uniformly bounded by a constant $\sigma \in \mathbb{R}_+$, such that for all k in $\mathbb{N}$, for all z in $\mathbb{R}^d$ we have: $\mathbb{E}[\|g_k(z) - \nabla F(z)\|^2] \leq \sigma^2$.

Theorem 1 (Convergence of compressed distributed SGD)

Under A1, A2, A3, if all $(F_i)_{i=1}^N$ are L-smooth, $\mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^N \mathcal{C}_{\mathrm{up}} \left(g_k^i (w_k) \right) \right)$ *is an unbiased stochastic oracle of $\nabla F(w_{k-1})$ with variance bounded by:*

$$\frac{2(\omega_{\mathrm{dwn}}+1)(\omega_{\mathrm{up}}+1)\sigma^2}{N} + \frac{4\omega_{\mathrm{dwn}}\omega_{\mathrm{up}}B^2}{N} + 2L\omega_{\mathrm{dwn}} \| w_k - w_* \|^2 \left(1 + \frac{2}{N}\right).$$

Compressed distributed SGD: $\forall k \in \mathbb{N}, w_k = w_{k-1} - \gamma \mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^N \mathcal{C}_{\mathrm{up}} (g_k^i (w_{k-1})) \right)$.



Relax the uniform bound

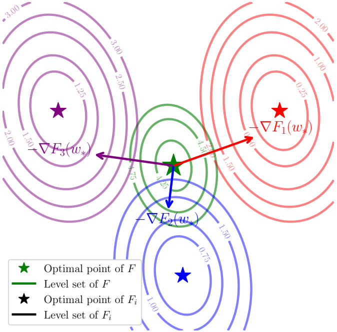
Remove the B^2-dependence

Theorem 1 (Convergence of compressed distributed SGD)

Under A1, A2, A3, if all $(F_i)_{i=1}^N$ are L-smooth, $\mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^N \mathcal{C}_{\mathrm{up}} (g_k^i (w_k)) \right)$ is an unbiased stochastic oracle of $\nabla F(w_{k-1})$ with variance bounded by:

$$\frac{2(\omega_{\mathrm{dwn}}+1)(\omega_{\mathrm{up}}+1)\sigma^2}{N} + \frac{4\omega_{\mathrm{dwn}}\omega_{\mathrm{up}}B^2}{N} + 2L\omega_{\mathrm{dwn}} \| w_k - w_* \|^2 \left(1 + \frac{2}{N}\right).$$

Compressed distributed SGD: $\forall k \in \mathbb{N}, w_k = w_{k-1} - \gamma \mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\mathrm{up}} (g_k^i (w_{k-1})) \right)$.



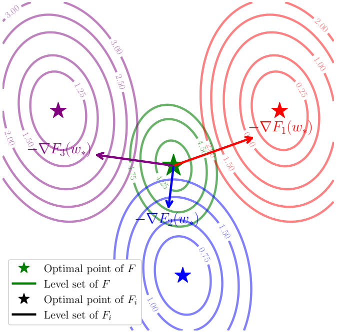
Remove the ω_{dwn}-dependence in the dominant term

Theorem 1 (Convergence of compressed distributed SGD)

Under A1, A2, A3, if all $(F_i)_{i=1}^N$ are L-smooth, $\mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\mathrm{up}} (g_k^i (w_k)) \right)$ is an unbiased stochastic oracle of $\nabla F(w_{k-1})$ with variance bounded by:

$$\frac{2(\omega_{\mathrm{dwn}} + 1)(\omega_{\mathrm{up}} + 1)\sigma^2}{N} + \frac{4\omega_{\mathrm{dwn}}\omega_{\mathrm{up}} B^2}{N} + 2L\omega_{\mathrm{dwn}} \| w_k - w_* \|^2 (1 + \frac{2}{N}).$$

Compressed distributed SGD: $\forall k \in \mathbb{N}, w_k = w_{k-1} - \gamma \mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^N \mathcal{C}_{\mathrm{up}} (g_k^i (w_{k-1})) \right)$.



Going beyond the worst-case assumption

Theorem 1 (Convergence of compressed distributed SGD)

Under A1, A2, A3, if all $(F_i)_{i=1}^N$ are L-smooth, $\mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^N \mathcal{C}_{\mathrm{up}} (g_k^i (w_k)) \right)$ is an unbiased stochastic oracle of $\nabla F(w_{k-1})$ with variance bounded by:

$$\frac{2(\omega_{\mathrm{dwn}}+1)(\omega_{\mathrm{up}}+1)\sigma^2}{N} + \frac{4\omega_{\mathrm{dwn}}\omega_{\mathrm{up}}B^2}{N} + 2L\omega_{\mathrm{dwn}} \| w_k - w_* \|^2 \left(1 + \frac{2}{N}\right).$$

Contributions

Outline: bibliography

I. *Artemis: tight convergence guarantees for bidirectional compression with heterogeneous clients*, **P** and Dieuleveut, under review at *Journal of Parallel and Distributed Computing*

II. *MCM: a preserved central model for faster bidirectional compression in distributed settings*, **P** and Dieuleveut, Neurips 2021

III. *Convergence rates for distributed, compressed and averaged least-squares regression: application to Federated Learning*, **P** and Dieuleveut, under review at *Journal of Machine Learning Research*

Outline: bibliography

I. *Artemis: tight convergence guarantees for bidirectional compression with heterogeneous clients*, **P** and Dieuleveut, under review at *Journal of Parallel and Distributed Computing*

II. *MCM: a preserved central model for faster bidirectional compression in distributed settings*, **P** and Dieuleveut, Neurips 2021

III. *Convergence rates for distributed, compressed and averaged least-squares regression: application to Federated Learning*, **P** and Dieuleveut, under review at *Journal of Machine Learning Research*

Table 1: Summary of contributions.

	Bi-compr.	Heterogeneity	LSR	
I.	✓	✓		Interaction between compression and heterogeneity
II.	✓		(✓)	Asympt. cancels impact of down compression
III.		(✓)	✓	Beyond worst-case analysis

Outline: bibliography

I. *Artemis: tight convergence guarantees for bidirectional compression with heterogeneous clients*, **P** and Dieuleveut, under review at *Journal of Parallel and Distributed Computing*

II. *MCM: a preserved central model for faster bidirectional compression in distributed settings*, **P** and Dieuleveut, Neurips 2021

III. *Convergence rates for distributed, compressed and averaged least-squares regression: application to Federated Learning*, **P** and Dieuleveut, under review at *Journal of Machine Learning Research*

Table 1: Summary of contributions.

	Bi-compr.	Heterogeneity	LSR	
I.	✓	✓		Interaction between compression and heterogeneity
II.	✓		(✓)	Asympt. cancels impact of down compression
III.		(✓)	✓	Beyond worst-case analysis

Not included in my manuscript: *FLamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings*, Ogier du Terrail, [...] **P**, [...] Andreux, Neurips 2022.

I. `Artemis` and the memory mechanism

Assumptions

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

Assumption 4 (Cocoercivity)

All $(g_k^i)_{i=1}^N$ stochastic gradient are L-cocoercive in quadratic mean.

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

Assumption 4 (Cocoercivity)

All $(g_k^i)_{i=1}^N$ *stochastic gradient are* L*-cocoercive in quadratic mean.*

Assumption 5 (Strong-convexity)

F *is strongly-convex.*

Assumptions

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

Assumption 4 (Cocoercivity)

All $(g_k^i)_{i=1}^N$ stochastic gradient are L-cocoercive in quadratic mean.

Assumption 5 (Strong-convexity)

F is strongly-convex.

Extension: We extend our results to the **convex case**.

Assumptions

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

Assumption 4 (Cocoercivity)

All $(g_k^i)_{i=1}^N$ stochastic gradient are L-cocoercive in quadratic mean.

Assumption 5 (Strong-convexity)

F is strongly-convex.

Extension: We extend our results to the **convex case**.

Assumption 6 (Noise over stochastic gradients computation)

The noise over stochastic gradients for a mini-batch of size b, is bounded at w_*:

$$\exists \sigma_* \in \mathbb{R}_+, \quad \forall k \in \mathbb{N}, \quad \forall i \in [\![1, N]\!], \quad \forall w \in \mathbb{R}^d : \qquad E[\| g_k^i(w_*) - \nabla F_i(w_*) \|^2] \leq \sigma_*^2 / b.$$

[As in GLQ+19, DDB20]

The memory mechanism to tackle heterogeneous clients

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

Consequence of clients' heterogeneity: $\lim_{k \to +\infty} g_{k+1}^i (w_*) \neq 0$.

The memory mechanism to tackle heterogeneous clients

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

Consequence of clients' heterogeneity: $\lim_{k \to +\infty} g_{k+1}^i (w_*) \neq 0$.

Goal: Compress a quantity that goes to 0

Solution: Compute (on the server and the worker independently) a **"memory"** h_k^i s.t.

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

The memory mechanism to tackle heterogeneous clients

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

Consequence of clients' heterogeneity: $\lim_{k \to +\infty} g_{k+1}^i (w_*) \neq 0$.

Goal: Compress a quantity that goes to 0

Solution: Compute (on the server and the worker independently) a **"memory"** h_k^i s.t.

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

⇒ The update equation becomes:

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

where α is the memory's learning rate.

The memory mechanism to tackle heterogeneous clients

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

Consequence of clients' heterogeneity: $\lim_{k \to +\infty} g_{k+1}^i (w_*) \neq 0.$

Goal: Compress a quantity that goes to 0

Solution: Compute (on the server and the worker independently) a **"memory"** h_k^i s.t.

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

$\Rightarrow$ The update equation becomes:

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

where α is the memory's learning rate.

$\Longrightarrow$ **Introducing this uplink memory mechanism is crucial to handle data heterogeneity, see Theorem 2.**

The memory mechanism to tackle heterogeneous clients

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

Consequence of clients' heterogeneity: $\lim_{k \to +\infty} g_{k+1}^i (w_*) \neq 0$.

Goal: Compress a quantity that goes to 0

Solution: Compute (on the server and the worker independently) a **"memory"** h_k^i s.t.

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

$\Rightarrow$ The update equation becomes:

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

γ: SGD step-size

α: memory's learning rate

where α is the memory's learning rate.

$\Longrightarrow$ **Introducing this uplink memory mechanism is crucial to handle data heterogeneity, see Theorem 2.**

Convergence for an L–smooth and μ–strongly convex F

Theorem 2 (Convergence of `Artemis`)

Under A1-2 and A4-6, for a step size γ under some conditions, for a learning rate α and for any k in $\mathbb{N}$,

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

with:

Variant	Var
$\alpha = 0$	$(\omega_{\mathrm{dwn}}+1)(\omega_{\mathrm{up}}+1)(\sigma_*^2 + B^2)$
$\alpha(\omega_{\mathrm{up}}+1) = 1/2$	$(\omega_{\mathrm{dwn}}+1)(\omega_{\mathrm{up}}+1)\sigma_*^2$

Convergence for an L–smooth and μ–strongly convex F

Theorem 2 (Convergence of `Artemis`)

Under A1-2 and A4-6, for a step size γ under some conditions, for a learning rate α and for any k in $\mathbb{N}$,

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

with:

Variant	Var
$\alpha = 0$	$(\omega_{\mathrm{dwn}}+1)(\omega_{\mathrm{up}}+1)(\sigma_*^2 + B^2)$
$\alpha(\omega_{\mathrm{up}}+1) = 1/2$	$(\omega_{\mathrm{dwn}}+1)(\omega_{\mathrm{up}}+1)\sigma_*^2$

- Linear rate up to a constant of the order of Var

Convergence for an L–smooth and μ–strongly convex F

Theorem 2 (Convergence of `Artemis`)

Under A1-2 and A4-6, for a step size γ under some conditions, for a learning rate α and for any k in $\mathbb{N}$,

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

with:

Variant	Var
$\alpha = 0$	$(\omega_{\mathrm{dwn}}+1)(\omega_{\mathrm{up}}+1)(\sigma_*^2 + B^2)$
$\alpha(\omega_{\mathrm{up}}+1) = 1/2$	$(\omega_{\mathrm{dwn}}+1)(\omega_{\mathrm{up}}+1)\sigma_*^2$

- Linear rate up to a constant of the order of Var
- The variance (Var) increases with the compression level.

Convergence for an $L-$smooth and $\mu-$strongly convex F

Theorem 2 (Convergence of `Artemis`)

Under A1-2 and A4-6, for a step size γ under some conditions, for a learning rate α and for any k in $\mathbb{N}$,

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

with:

Variant	Var
$\alpha = 0$	$(\omega_{\mathrm{dwn}}+1)(\omega_{\mathrm{up}}+1)(\sigma_*^2 + B^2)$
$\alpha(\omega_{\mathrm{up}}+1) = 1/2$	$(\omega_{\mathrm{dwn}}+1)(\omega_{\mathrm{up}}+1)\sigma_*^2$

- Linear rate up to a constant of the order of Var
- The variance (Var) increases with the compression level.
- When $B^2 \neq 0$ (non-i.i.d. settings), if $\sigma_*^2 = 0$, then using memory ($\alpha \neq 0$) leads to linear convergence

Convergence for an L–smooth and μ–strongly convex F

Theorem 2 (Convergence of `Artemis`)

Under A1-2 and A4-6, for a step size γ under some conditions, for a learning rate α and for any k in $\mathbb{N}$,

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

with:

Variant	Var
$\alpha = 0$	$(\omega_{\mathrm{dwn}} + 1)(\omega_{\mathrm{up}} + 1)(\sigma_*^2 + B^2)$
$\alpha(\omega_{\mathrm{up}} + 1) = 1/2$	$(\omega_{\mathrm{dwn}} + 1)(\omega_{\mathrm{up}} + 1)\sigma_*^2$

- Linear rate up to a constant of the order of Var
- The variance (Var) increases with the compression level.
- When $B^2 \neq 0$ (non-i.i.d. settings), if $\sigma_*^2 = 0$, then using memory ($\alpha \neq 0$) leads to linear convergence
- If $B^2 = 0$ (i.i.d. settings), the memory is useless

Theorem 2 (Convergence of `Artemis`)

Under A1-2 and A4-6, for a step size γ under some conditions, for a learning rate α and for any k in $\mathbb{N}$,

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

with:

Variant	Var
$\alpha = 0$	$(\omega_{\mathrm{dwn}}+1)(\omega_{\mathrm{up}}+1)(\sigma_*^2 + B^2)$
$\alpha(\omega_{\mathrm{up}}+1) = 1/2$	$(\omega_{\mathrm{dwn}}+1)(\omega_{\mathrm{up}}+1)\sigma_*^2$

- Linear rate up to a constant of the order of Var
- The variance (Var) increases with the compression level.
- When $B^2 \neq 0$ (non-i.i.d. settings), if $\sigma_*^2 = 0$, then using memory ($\alpha \neq 0$) leads to linear convergence
- If $B^2 = 0$ (i.i.d. settings), the memory is useless
- Recovers classical SGD rate in the absence of compression

Convergence for an $L-$smooth and $\mu-$strongly convex F

Theorem 2 (Convergence of `Artemis`)

Under A1-2 and A4-6, for a step size γ under some conditions, for a learning rate α and for any k in $\mathbb{N}$,

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

with:

Variant	Var
$\alpha = 0$	$(\omega_{\mathrm{dwn}}+1)(\omega_{\mathrm{up}}+1)(\sigma_*^2 + B^2)$
$\alpha(\omega_{\mathrm{up}}+1) = 1/2$	$(\omega_{\mathrm{dwn}}+1)(\omega_{\mathrm{up}}+1)\sigma_*^2$

- Linear rate up to a constant of the order of Var
- The variance (Var) increases with the compression level.
- When $B^2 \neq 0$ (non-i.i.d. settings), if $\sigma_*^2 = 0$, then using memory ($\alpha \neq 0$) leads to linear convergence
- If $B^2 = 0$ (i.i.d. settings), the memory is useless
- Recovers classical SGD rate in the absence of compression

Theorem 3 (Lower bound on the variance for *linear* compressors)

Under A1-2 and A4-6, for $\gamma, \alpha_{\mathrm{up}}, E$ given in Theorem 2, for Θ_k the distribution of w_k.

There exists a limit distribution $\pi_{\gamma,\alpha}$ s.t. for any $k \geq 1$, for C_0 a constant:

$$\mathcal{W}_2(\Theta_k, \pi_{\gamma,v}) \leq (1-\gamma\mu)^k C_0 .$$

Furthermore:

$$\mathbb{E}[\| w_k - w_* \|^2] \xrightarrow[k\to\infty]{} \mathbb{E}_{w\sim\pi_{\gamma,v}}[\| w - w_* \|^2]$$

which is lower bounded s.t.:

$$\mathbb{E}_{w\sim\pi_{\gamma,v}}[\| w - w_* \|^2] \underset{\gamma\to 0}{=} \Omega(\gamma \mathrm{Var}/\mu N) .$$

Theorem 2 (Convergence of Artemis)

Under A1-2 and A4-6, for a step size γ under some conditions, for a learning rate α and for any k in $\mathbb{N}$,

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

with:

Variant	Var
$\alpha = 0$	$(\omega_{\mathrm{dwn}}+1)(\omega_{\mathrm{up}}+1)(\sigma_*^2 + B^2)$
$\alpha(\omega_{\mathrm{up}}+1) = 1/2$	$(\omega_{\mathrm{dwn}}+1)(\omega_{\mathrm{up}}+1)\sigma_*^2$

- Linear rate up to a constant of the order of Var
- The variance (Var) increases with the compression level.
- When $B^2 \neq 0$ (non-i.i.d. settings), if $\sigma_*^2 = 0$, then using memory ($\alpha \neq 0$) leads to linear convergence
- If $B^2 = 0$ (i.i.d. settings), the memory is useless
- Recovers classical SGD rate in the absence of compression

Theorem 3 (Lower bound on the variance for *linear* compressors)

Under A1-2 and A4-6, for $\gamma, \alpha_{\mathrm{up}}, E$ given in Theorem 2, for Θ_k the distribution of w_k.

There exists a limit distribution $\pi_{\gamma,\alpha}$ s.t. for any $k \geq 1$, for C_0 a constant:

$$\mathcal{W}_2(\Theta_k, \pi_{\gamma,v}) \leq (1-\gamma\mu)^k C_0 .$$

Furthermore:

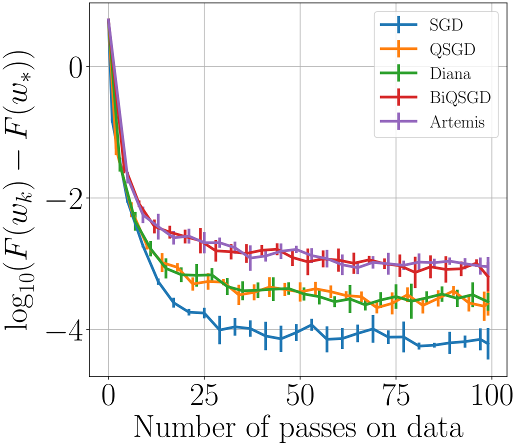
$$\mathbb{E}[\|w_k - w_*\|^2] \xrightarrow[k\to\infty]{} \mathbb{E}_{w\sim\pi_{\gamma,v}}[\|w - w_*\|^2]$$

which is lower bounded s.t.:

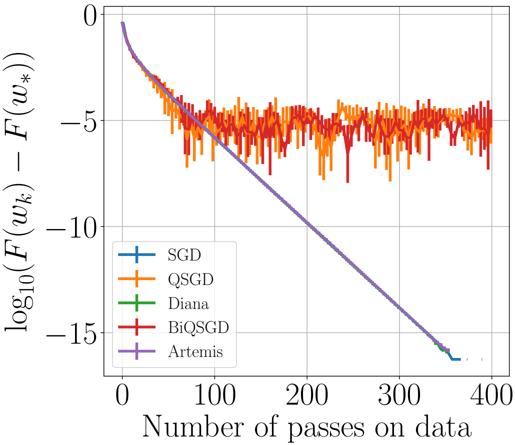
$$\mathbb{E}_{w\sim\pi_{\gamma,v}}[\|w - w_*\|^2] \underset{\gamma\to 0}{=} \Omega(\gamma \mathrm{Var}/\mu N) .$$

The quadratic increase in the variance is not an artifact of the proof!

- Left: illustration of the saturation when $\sigma_*^2 \neq 0$ and data is i.i.d.
- Right: illustration of the memory benefits when $\sigma_*^2 = 0$ but with non-i.i.d. data.



(a) Least-square reg. (i.i.d.): $\sigma_*^2 \neq 0$



(b) Logisitc reg. (non-i.i.d.): $\sigma_*^2 = 0$.

Figure 4: Synthetic datasets

- Left: almost homogeneous clients.
- Right: heterogeneous clients.
- Stochastic gradient descent: $\sigma_* \neq 0$.

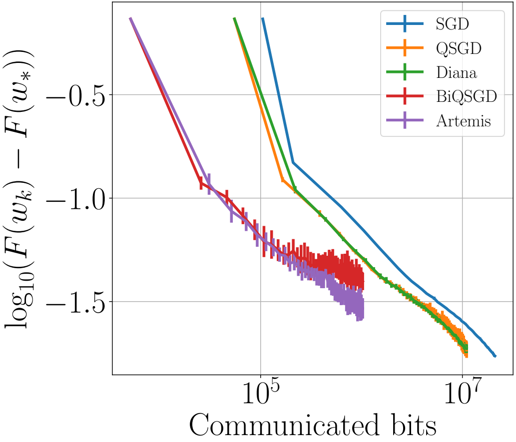


Figure 5: Superconduct (LSR), $b = 64$

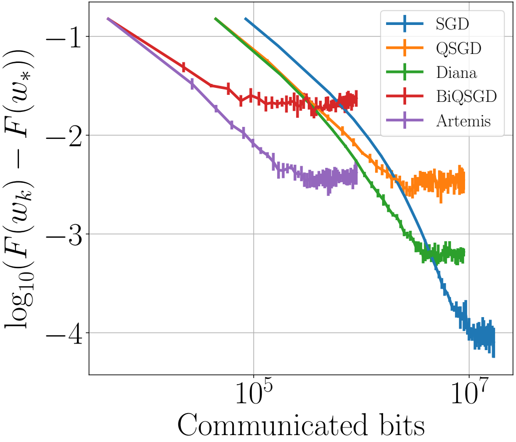


Figure 6: Quantum (LR), $b = 256$

Partial conclusion

Take-away 1

- ***Bidirectional compression*** *to reduce the communication cost.*

Partial conclusion

Take-away 1

- ***Bidirectional compression*** *to reduce the communication cost.*

Take-away 2

- *Primary factor:* ***noise*** σ_* ***on the gradient*** *computed on the* ***optimal point****.*
- *Key impact of* ***memory*** *on* ***non-i.i.d. data****.*

Partial conclusion

Take-away 1

- ***Bidirectional compression*** *to reduce the communication cost.*

Take-away 2

- *Primary factor:* ***noise*** σ_* ***on the gradient*** *computed on the* ***optimal point****.*
- *Key impact of* ***memory*** *on* ***non-i.i.d. data****.*

Take-away 3

- ***Lower bound*** *on the asymptotic variance.*

II. MCM and the preserved update equation

Classical approach - **degrade the model on the central server.**

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

The gradient is taken at the point w_k held by the central server [LLTY20, PD20, TYL+19, ZHK19].

Classical approach vs new approach

Classical approach - **degrade the model on the central server.**

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

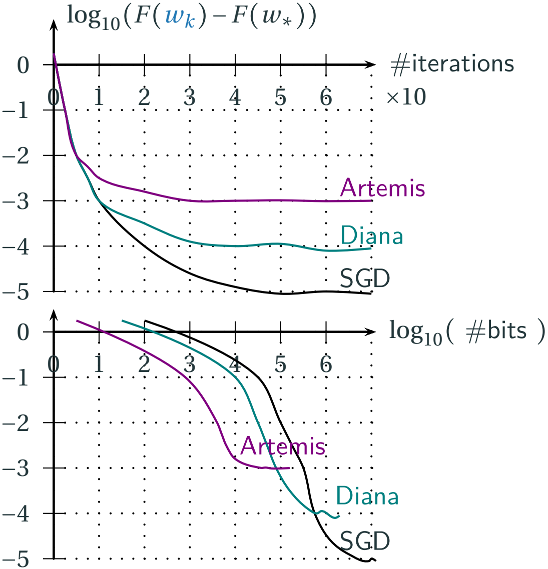
The gradient is taken at the point w_k held by the central server [LLTY20, PD20, TYL$^+$19, ZHK19].

New approach - **preserve the model on the central server.**

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

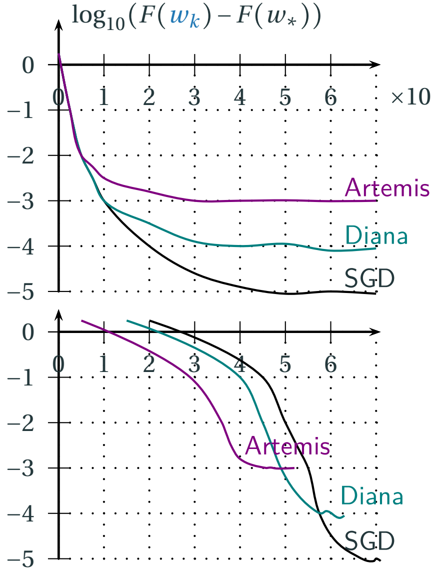
The gradient is taken at a random point $\hat{w}_k$ s.t. $\mathbb{E}[\hat{w}_k | w_k] = w_k$.

Classical approach

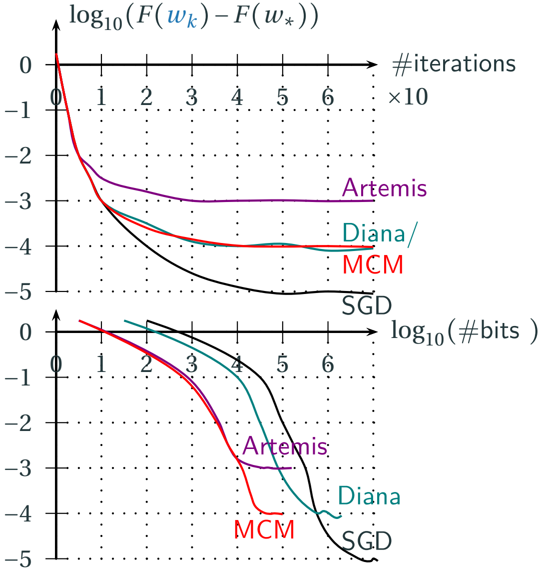


What do we hope for? (using a constant step-size γ)

Classical approach



New approach



The downlink memory mechanism for `MCM`

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

1. available on both clients and central server
2. the difference Ω_k between the model and this memory is compressed and exchanged
3. the local model is reconstructed from this information

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

The downlink memory mechanism for MCM

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

1. available on both clients and central server
2. the difference Ω_k between the model and this memory is compressed and exchanged
3. the local model is reconstructed from this information

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

$\Longrightarrow$Introducing this memory mechanism is crucial to control the variance of the local model $\hat{w}_k$.

The downlink memory mechanism for MCM

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

1. available on both clients and central server
2. the difference Ω_k between the model and this memory is compressed and exchanged
3. the local model is reconstructed from this information

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

$\Longrightarrow$Introducing this memory mechanism is crucial to control the variance of the local model $\hat{w}_k$.

⚠We still use the *uplink memory term* (required to tackle the heterogeneous settings). ⚠

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

1. available on both clients and central server
2. the difference Ω_k between the model and this memory is compressed and exchanged
3. the local model is reconstructed from this information

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

$\Longrightarrow$**Introducing this memory mechanism is crucial to control the variance of the local model $\hat{w}_k$.**

⚠We still use the *uplink memory term* (required to tackle the heterogeneous settings). ⚠

$\Longrightarrow$ **This is MCM.**

Convergence theorem

Assumption 7 (Smoothness and convexity.)

F is convex, twice continuously differentiable and L-smooth.

Convergence theorem

Assumption 7 (Smoothness and convexity.)

F is convex, twice continuously differentiable and L-smooth.

Theorem 4 (Convergence of `MCM`, convex case)

Under A1, A3, A7, for K in $\mathbb{N}$, with a large enough step-size $\gamma = \sqrt{\frac{\delta_0^2 N b}{(\omega_{\mathrm{up}}+1)\sigma^2 K}}$, denoting $\bar{w}_K = \frac{1}{K}\sum_{i=0}^{K-1} w_i$, we have:

$$\mathbb{E}\left[F(\bar{w}_K) - F_*\right] \leq 2\sqrt{\frac{\delta_0^2(\omega_{\mathrm{up}}+1)\sigma^2}{NbK}} + O\left(\frac{\omega_{\mathrm{up}}\omega_{\mathrm{dwn}}}{K}\right).$$

Convergence theorem

Assumption 7 (Smoothness and convexity.)

F is convex, twice continuously differentiable and L-smooth.

Theorem 4 (Convergence of `MCM`, convex case)

Under A1, A3, A7, for K in $\mathbb{N}$, with a large enough step-size $\gamma = \sqrt{\frac{\delta_0^2 N b}{(\omega_{\mathrm{up}}+1)\sigma^2 K}}$, denoting $\bar{w}_K = \frac{1}{K}\sum_{i=0}^{K-1} w_i$, we have:

$$\mathbb{E}\left[F(\bar{w}_K) - F_*\right] \leq \underbrace{2\sqrt{\frac{\delta_0^2(\omega_{\mathrm{up}}+1)\sigma^2}{NbK}}}_{\text{dominant term}} + \underbrace{O\left(\frac{\omega_{\mathrm{up}}\omega_{\mathrm{dwn}}}{K}\right)}_{\text{lower order term}}.$$

Convergence theorem

Assumption 7 (Smoothness and convexity.)

F is convex, twice continuously differentiable and L-smooth.

Theorem 4 (Convergence of `MCM`, convex case)

Under A1, A3, A7, for K in $\mathbb{N}$, with a large enough step-size $\gamma = \sqrt{\frac{\delta_0^2 N b}{(\omega_{\text{up}}+1)\sigma^2 K}}$, denoting $\bar{w}_K = \frac{1}{K}\sum_{i=0}^{K-1} w_i$, we have:

$$\mathbb{E}\left[F(\bar{w}_K) - F_*\right] \leq \underbrace{2\sqrt{\frac{\delta_0^2(\omega_{\text{up}}+1)\sigma^2}{NbK}}}_{\text{dominant term}} + \underbrace{O\left(\frac{\omega_{\text{up}}\omega_{\text{dwn}}}{K}\right)}_{\text{lower order term}} .$$

- independent of ω_{dwn} → dominant term
- identical to `Diana` (uni-compression)

- depends on ω_{dwn} → lower order term
- asymptotically negligible

Convergence theorem

Assumption 7 (Smoothness and convexity.)

F is convex, twice continuously differentiable and L-smooth.

Theorem 4 (Convergence of MCM, convex case)

Under A1, A3, A7, for K in $\mathbb{N}$, with a large enough step-size $\gamma = \sqrt{\frac{\delta_0^2 N b}{(\omega_{\mathrm{up}}+1)\sigma^2 K}}$, denoting $\bar{w}_K = \frac{1}{K}\sum_{i=0}^{K-1} w_i$, we have:

$$\mathbb{E}\left[F(\bar{w}_K) - F_*\right] \leq 2\sqrt{\frac{\delta_0^2(\omega_{\mathrm{up}}+1)\sigma^2}{NbK}} + O\left(\frac{\omega_{\mathrm{up}}\omega_{\mathrm{dwn}}}{K}\right).$$

Moreover if $\sigma^2 = 0$, we recover a faster convergence:

$$\mathbb{E}[F(\bar{w}_K) - F_*] = O(K^{-1}) \ .$$

Assumption 7 (Smoothness and convexity.)

F is convex, twice continuously differentiable and L-smooth.

Theorem 4 (Convergence of MCM, convex case)

Under A1, A3, A7, for K in $\mathbb{N}$, with a large enough step-size $\gamma = \sqrt{\frac{\delta_0^2 N b}{(\omega_{\mathrm{up}}+1)\sigma^2 K}}$, denoting $\bar{w}_K = \frac{1}{K}\sum_{i=0}^{K-1} w_i$, we have:

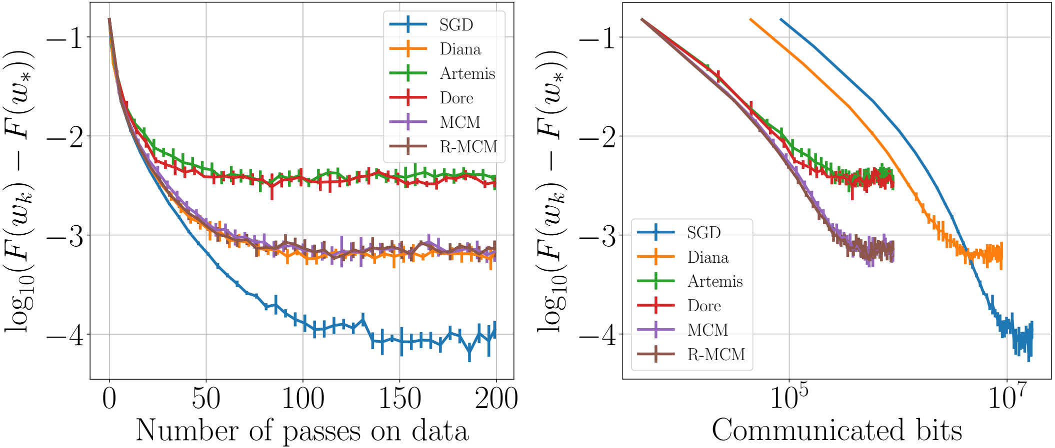
$$\mathbb{E}\left[F(\bar{w}_K) - F_*\right] \leq 2\sqrt{\frac{\delta_0^2(\omega_{\mathrm{up}}+1)\sigma^2}{NbK}} + O\left(\frac{\omega_{\mathrm{up}}\omega_{\mathrm{dwn}}}{K}\right).$$

Moreover if $\sigma^2 = 0$, we recover a faster convergence:

$$\mathbb{E}[F(\bar{w}_K) - F_*] = O(K^{-1})\ .$$

Remark: this result is also extended to both strongly-convex and non-convex cases.

Experiments in convex settings (using a constant step-size γ)



(a) X axis in # iterations

(b) X axis in # bits

Figure 7: Quantum with $b = 400$, $\gamma = 1/L$ (Logistic regression).

Experiments in non-convex settings

Nonconvex framework	MNIST (CNN, d=2e4, 4 bits-quantization with norm 2)	Fashion MNIST (FashionSimpleNet, d=4e5, 4 bits-quantization with norm 2)	Heterogeneous EMNIST (CNN, d=2e4, 4 bits-quantization with norm 2)	CIFAR-10 (LeNet, d=62e3, 16 bits-quantization with norm 2)
Accuracy after 300 epochs	SGD: 99.0%	SGD: 92.4%	SGD: 99.0%	SGD: 69.1%
	Diana: 98.9% MCM: 98.8%	Diana: 92.4% MCM: 90.6%	Diana: 98.9% MCM: 98.9%	Diana: 64.0% MCM: 63.5%
	Artemis: 97.9% Dore: 97.9%	Artemis: 86.7% Dore: 87.9%	Artemis: 98.3% Dore: 98.5%	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	Diana: 0.141 MCM: 0.209	Diana: 0.031 MCM: 0.030	Diana: 1.047 MCM: 1.096
	Artemis: 0.075 Dore: 0.072	Artemis: 0.332 Dore: 0.300	Artemis: 0.052 Dore: 0.048	Artemis: 1.342 Dore: 1.292

Partial conclusion

Take-away 4

- *New algorithm to perform **bidirectional compression**.*
- *Asymptotically same rate of convergence than **unidirectional compression**.*

Take-away 5

- *Local gradients computed on a **"perturbed model"** (more challenging).*

Additional contributions of the article:

- Randomized-`MCM` with independent compressions: improves convergence in the quadratic case.

III. Beyond worst-case analysis

Back to the URBV assumption

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

Big question: what is the impact of $\mathcal{C}$ on convergence?

↬ To limit the number of bits exchanged, we **compress** the uplink signal before transmitting it.
Big question: what is the impact of $\mathcal{C}$ on convergence?
Compressed distributed SGD: $\forall k \in \mathbb{N}, w_k = w_{k-1} - \frac{\gamma}{N} \sum_{i=1}^{N} \mathcal{C}(g_k^i(w_{k-1}))$.

Assumption

There exists a constant $\omega \in \mathbb{R}_+^*$ s.t. $\mathcal{C}$ satisfies, for all z in $\mathbb{R}^d$:

$$\mathbb{E}[\mathcal{C}(z)] = z \quad \text{and} \quad \mathbb{E}\left[\|\mathcal{C}(z) - z\|^2\right] \leq \omega \|z\|^2 .$$

Back to the URBV assumption

→ To limit the number of bits exchanged, we **compress** the uplink signal before transmitting it.
Big question: what is the impact of $\mathcal{C}$ on convergence?
Compressed distributed SGD: $\forall k \in \mathbb{N}, w_k = w_{k-1} - \frac{\gamma}{N} \sum_{i=1}^{N} \mathcal{C}(g_k^i(w_{k-1}))$.

Assumption

There exists a constant $\omega \in \mathbb{R}_+^*$ s.t. $\mathcal{C}$ satisfies, for all z in $\mathbb{R}^d$:

$$\mathbb{E}[\mathcal{C}(z)] = z \quad \textit{and} \quad \mathbb{E}\left[\|\mathcal{C}(z) - z\|^2\right] \leq \omega \|z\|^2 .$$

- To go beyond this *worst-case* assumption and provide a tighter analyse.
- Focus on the LSR framework, which is popular for fine-grained analyses.

Back to the URBV assumption

↬ To limit the number of bits exchanged, we **compress** the uplink signal before transmitting it.
Big question: what is the impact of $\mathcal{C}$ on convergence?
Compressed distributed SGD: $\forall k \in \mathbb{N}, w_k = w_{k-1} - \frac{\gamma}{N} \sum_{i=1}^{N} \mathcal{C}(g_k^i(w_{k-1}))$.

Assumption

There exists a constant $\omega \in \mathbb{R}_+^*$ s.t. $\mathcal{C}$ satisfies, for all z in $\mathbb{R}^d$:

$$\mathbb{E}[\mathcal{C}(z)] = z \quad \textit{and} \quad \mathbb{E}\left[\|\mathcal{C}(z) - z\|^2\right] \leq \omega \|z\|^2 .$$

- To go beyond this *worst-case* assumption and provide a tighter analyse.
- Focus on the LSR framework, which is popular for fine-grained analyses.

Final goal: highlight the differences in convergence between several unbiased compression schemes having the *same* variance increase.

Back to the URBV assumption

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

Big question: what is the impact of $\mathcal{C}$ on convergence?

Compressed distributed SGD: $\forall k \in \mathbb{N}, w_k = w_{k-1} - \frac{\gamma}{N} \sum_{i=1}^N \mathcal{C}(g_k^i(w_{k-1}))$.

Assumption

There exists a constant $\omega \in \mathbb{R}_+^*$ s.t. $\mathcal{C}$ satisfies, for all z in $\mathbb{R}^d$:

$$\mathbb{E}[\mathcal{C}(z)] = z \quad \text{and} \quad \mathbb{E}\left[\|\mathcal{C}(z) - z\|^2\right] \leq \omega \|z\|^2 .$$

- To go beyond this *worst-case* assumption and provide a tighter analyse.
- Focus on the LSR framework, which is popular for fine-grained analyses.

Simplified setting for this presentation:

- $N = 1$ client.
- The client accesses K in $\mathbb{N}^*$ i.i.d. observations $(x_k, y_k)_{k \in \{1,\dots,K\}} \sim \mathcal{D}^{\otimes K}$, such that there exists a well-defined model w_*:

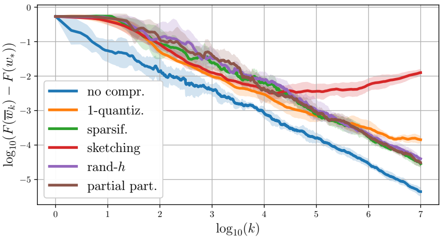
$$\forall k \in \{1, \dots, K\}, \quad y_k = \langle x_k, w_* \rangle + \varepsilon_k^i, \qquad \text{with} \quad \varepsilon_k \sim \mathcal{N}(0, \sigma^2) \ .$$

Comparing various compressors in different scenarios

5 compressors: 4 scenarios, 4 different behaviors.

Comparing various compressors in different scenarios

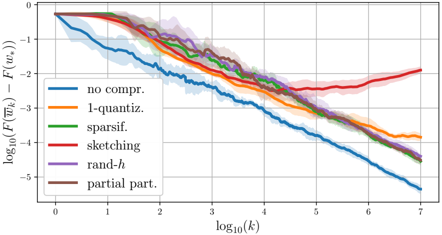
5 compressors: 4 scenarios, 4 different behaviors.



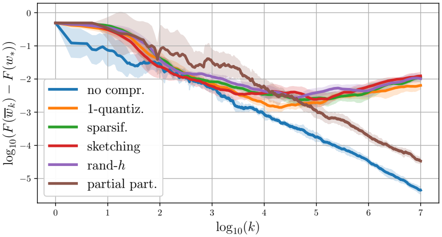
Sketching is very bad, quantiz. is slightly worse.

Comparing various compressors in different scenarios

5 compressors: 4 scenarios, 4 different behaviors.



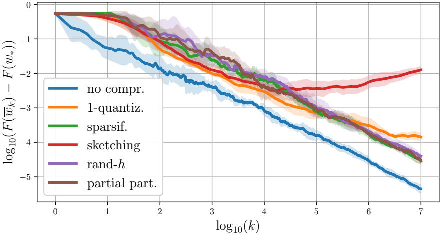
Sketching is very bad, quantiz. is slightly worse.



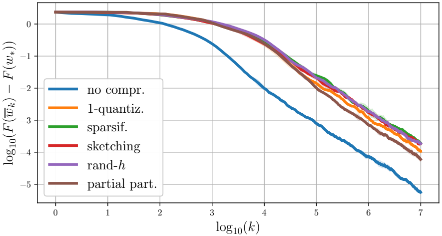
Only partial part. is good.

Comparing various compressors in different scenarios

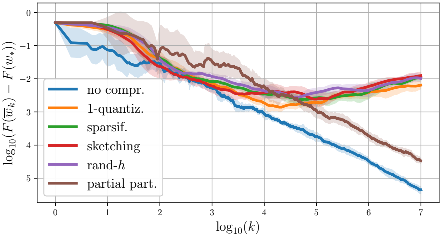
5 compressors: 4 scenarios, 4 different behaviors.



Sketching is very bad, quantiz. is slightly worse.



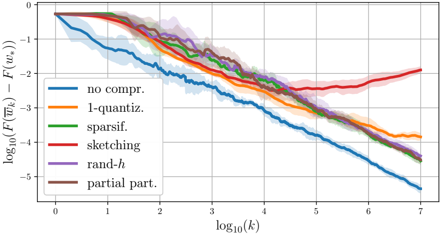
All compressors are equivalent and behave well.



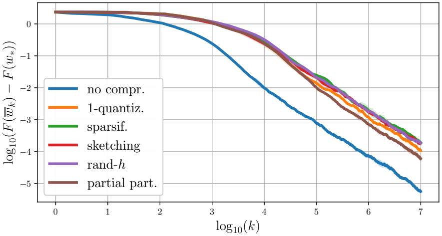
Only partial part. is good.

Comparing various compressors in different scenarios

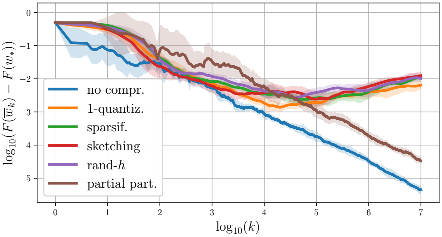
5 compressors: 4 scenarios, 4 different behaviors.



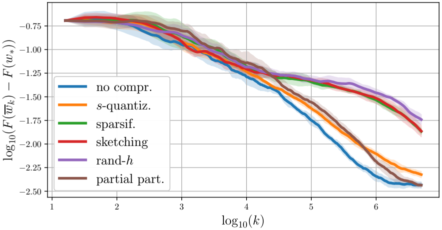
Sketching is very bad, quantiz. is slightly worse.



All compressors are equivalent and behave well.



Only partial part. is good.

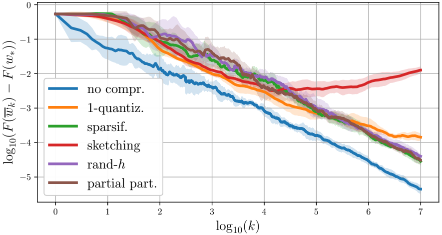


Quantiz. and partial part. are good.

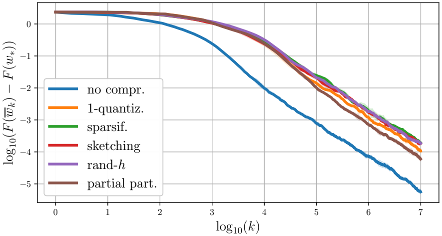
Comparing various compressors in different scenarios

5 compressors: 4 scenarios, 4 different behaviors.

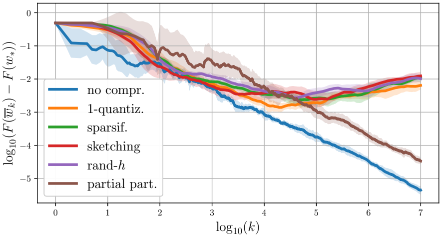
Can we explain this four different behaviors?



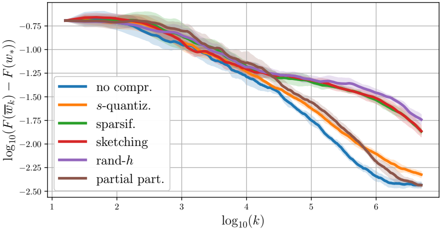
Sketching is very bad, quantiz. is slightly worse.



All compressors are equivalent and behave well.



Only partial part. is good.



Quantiz. and partial part. are good.

Definition 1 (Linear Stochastic Approximation, LSA)

Let $w_0 \in \mathbb{R}^d$ be the initialization, the linear stochastic approximation[1] recursion is defined as:

$$w_k = w_{k-1} - \gamma \nabla F(w_{k-1}) + \gamma \xi_k(w_{k-1} - w_*), \quad k \in \mathbb{N}, \tag{LSA}$$

- $\gamma > 0$: *step size,*
- $(\xi_k)_{k \in \mathbb{N}^*}$: *sequence of i.i.d. zero-centered random fields that characterizes the stochastic oracle on $\nabla F(\cdot)$.*

[1] While in LSA literature, both the mean-field ∇F and the noise-field (ξ_k) are linear, we do not here consider the noise fields to be linear.

Definition 1 (Linear Stochastic Approximation, LSA)

Let $w_0 \in \mathbb{R}^d$ be the initialization, the linear stochastic approximation[1] recursion is defined as:

$$w_k = w_{k-1} - \gamma \nabla F(w_{k-1}) + \gamma \xi_k (w_{k-1} - w_*), \quad k \in \mathbb{N}, \tag{LSA}$$

- $\gamma > 0$: *step size,*
- $(\xi_k)_{k \in \mathbb{N}^*}$: *sequence of i.i.d. zero-centered random fields that characterizes the stochastic oracle on* $\nabla F(\cdot)$.

We assume F quadratic:

- H_F: *its Hessian*
- μ: *its smallest eigenvalue.*

For any k in $\mathbb{N}$, with $\eta_k = w_k - w_*$, we get equivalently:

$$\eta_k = (\mathrm{I} - \gamma H_F)\eta_{k-1} + \gamma \xi_k (\eta_{k-1}), \quad k \in \mathbb{N}.$$

[1] While in LSA literature, both the mean-field ∇F and the noise-field (ξ_k) are linear, we do not here consider the noise fields to be linear.

Algorithm 1 (LMS with a single worker)

We have for all $k \in \mathbb{N}$:

$$w_k = w_{k-1} - \gamma(\langle w_{k-1}, x_k\rangle - y_k)x_k\,,$$

Equivalently, for $w \in \mathbb{R}^d$:

$$\xi_k(w) = (x_k x_k^\top - \mathbb{E}[x_1 x_1^\top])w + (\langle w_*, x_k\rangle - y_k)x_k\,.$$

Examples and challenge

Algorithm 1 (LMS with a single worker)

We have for all $k \in \mathbb{N}$:

$$w_k = w_{k-1} - \gamma(\langle w_{k-1}, x_k \rangle - y_k) x_k ,$$

Equivalently, for $w \in \mathbb{R}^d$:

$$\xi_k(w) = (x_k x_k^\top - \mathbb{E}[x_1 x_1^\top]) w + (\langle w_*, x_k \rangle - y_k) x_k .$$

Algorithm 2 (Centralized compressed LMS)

At any step k in $\{1, \ldots, K\}$, we have an oracle $g_k(\cdot)$ of the gradient of the objective function F and a random compression mechanism $\mathcal{C}_k(\cdot)$.

For any step-size $\gamma > 0$ and any $k \in \mathbb{N}^*$, the resulting sequence of iterates $(w_k)_{k \in \mathbb{N}}$ satisfies:

$$w_k = w_{k-1} - \gamma \mathcal{C}_k(g_k(w_{k-1})) .$$

Equivalently, for $w \in \mathbb{R}^d$:

$$\xi_k(w) = \nabla F(w) - \mathcal{C}_k(g_k(w)) .$$

Algorithm 1 (LMS with a single worker)

We have for all $k \in \mathbb{N}$:

$$w_k = w_{k-1} - \gamma(\langle w_{k-1}, x_k \rangle - y_k) x_k ,$$

Equivalently, for $w \in \mathbb{R}^d$:

$$\xi_k(w) = (x_k x_k^\top - \mathbb{E}[x_1 x_1^\top]) w + (\langle w_*, x_k \rangle - y_k) x_k .$$

Algorithm 2 (Centralized compressed LMS)

At any step k in $\{1, \dots, K\}$, we have an oracle $g_k(\cdot)$ of the gradient of the objective function F and a random compression mechanism $\mathcal{C}_k(\cdot)$.

For any step-size $\gamma > 0$ and any $k \in \mathbb{N}^*$, the resulting sequence of iterates $(w_k)_{k \in \mathbb{N}}$ satisfies:

$$w_k = w_{k-1} - \gamma \mathcal{C}_k(g_k(w_{k-1})) .$$

Equivalently, for $w \in \mathbb{R}^d$:

$$\xi_k(w) = \nabla F(w) - \mathcal{C}_k(g_k(w)) .$$

Most analyses of (LSA) [Blu54, Lju77, LS83] assume either:

1. The field ξ_k is either linear [see KT03, BMP12, LP21] i.e. for any $z, z' \in \mathbb{R}^d$,

$$\xi_k(z) - \xi_k(z') = \xi_k(z - z') .$$

2. The noise-field is Lipschitz in squared expectation [MB11, Bac14, DDB20, GP23]. i.e. for any $z, z' \in \mathbb{R}^d$

$$\mathbb{E}[\|\xi_k(z) - \xi_k(z')\|^2] \leq C \|z - z'\|^2 .$$

Algorithm 1 (LMS with a single worker)

We have for all $k \in \mathbb{N}$:

$$w_k = w_{k-1} - \gamma(\langle w_{k-1}, x_k \rangle - y_k) x_k ,$$

Equivalently, for $w \in \mathbb{R}^d$:

$$\xi_k(w) = (x_k x_k^\top - \mathbb{E}[x_1 x_1^\top]) w + (\langle w_*, x_k \rangle - y_k) x_k .$$

Algorithm 2 (Centralized compressed LMS)

At any step k in $\{1, \dots, K\}$, we have an oracle $g_k(\cdot)$ of the gradient of the objective function F and a random compression mechanism $\mathcal{C}_k(\cdot)$.

For any step-size $\gamma > 0$ and any $k \in \mathbb{N}^*$, the resulting sequence of iterates $(w_k)_{k \in \mathbb{N}}$ satisfies:

$$w_k = w_{k-1} - \gamma \mathcal{C}_k(g_k(w_{k-1})) .$$

Equivalently, for $w \in \mathbb{R}^d$:

$$\xi_k(w) = \nabla F(w) - \mathcal{C}_k(g_k(w)) .$$

Most analyses of (LSA) [Blu54, Lju77, LS83] assume either:

1. The field ξ_k is either linear [see KT03, BMP12, LP21] i.e. for any $z, z' \in \mathbb{R}^d$,

$$\xi_k(z) - \xi_k(z') = \xi_k(z - z') .$$

2. The noise-field is Lipschitz in squared expectation [MB11, Bac14, DDB20, GP23]. i.e. for any $z, z' \in \mathbb{R}^d$

$$\mathbb{E}[\|\xi_k(z) - \xi_k(z')\|^2] \le C \|z - z'\|^2 .$$

$\Longrightarrow$ Specificity and bottleneck of compression: the resulting field **does not** satisfy such assumptions.

Definition 2 (Additive and multiplicative noise)

Under the setting of (LSA), for any k in $\mathbb{N}^*$:

$$\xi_k^{\text{add}} := \xi_k(0) \quad \textit{and} \quad \xi_k^{\text{mult}} : z \in \mathbb{R}^d \mapsto \xi_k(z) - \xi_k^{\text{add}}.$$

Definition 2 (Additive and multiplicative noise)

Under the setting of (LSA), for any k in $\mathbb{N}^*$:

$$\xi_k^{\text{add}} := \xi_k(0) \quad \textit{and} \quad \xi_k^{\text{mult}} : z \in \mathbb{R}^d \mapsto \xi_k(z) - \xi_k^{\text{add}}.$$

Assumption (Second moment of the multiplicative noise)

$\exists \mathcal{M}_1, \mathcal{M}_2 > 0$ *s.t. for any* η *in* $\mathbb{R}^d$:

1. $\mathbb{E}[\|\xi_1^{\text{mult}}(\eta)\|^2] \leq 2\mathcal{M}_2 \|H_F^{1/2}\eta\|^2 + 4\mathcal{A}.$
2. $\mathbb{E}[\|\xi_1^{\text{mult}}(\eta)\|^2] \leq \mathcal{M}_1 \|H_F^{1/2}\eta\| + 3\mathcal{M}_2 \|H_F^{1/2}\eta\|^2.$

Definition of $\mathfrak{C}_{\text{ania}}$

Definition 2 (Additive and multiplicative noise)

Under the setting of (LSA), for any k in $\mathbb{N}^*$:

$$\xi_k^{\text{add}} := \xi_k(0) \quad \textit{and} \quad \xi_k^{\text{mult}} : z \in \mathbb{R}^d \mapsto \xi_k(z) - \xi_k^{\text{add}}.$$

Assumption (Second moment of the multiplicative noise)

$\exists \mathcal{M}_1, \mathcal{M}_2 > 0$ *s.t. for any* η *in* $\mathbb{R}^d$:

1. $\mathbb{E}[\|\xi_1^{\text{mult}}(\eta)\|^2] \leq 2\mathcal{M}_2 \|H_F^{1/2}\eta\|^2 + 4\mathcal{A}.$
2. $\mathbb{E}[\|\xi_1^{\text{mult}}(\eta)\|^2] \leq \mathcal{M}_1 \|H_F^{1/2}\eta\| + 3\mathcal{M}_2 \|H_F^{1/2}\eta\|^2.$

Classical assumption

Hölder-type assumption (new)

Definition 2 (Additive and multiplicative noise)

Under the setting of (LSA), for any k in $\mathbb{N}^*$:

$$\xi_k^{\text{add}} := \xi_k(0) \quad \text{and} \quad \xi_k^{\text{mult}} : z \in \mathbb{R}^d \mapsto \xi_k(z) - \xi_k^{\text{add}}.$$

Assumption (Second moment of the multiplicative noise)

$\exists \mathcal{M}_1, \mathcal{M}_2 > 0$ *s.t. for any* η *in* $\mathbb{R}^d$:

1. $\mathbb{E}[\|\xi_1^{\text{mult}}(\eta)\|^2] \leq 2\mathcal{M}_2 \|H_F^{1/2}\eta\|^2 + 4\mathcal{A}.$
2. $\mathbb{E}[\|\xi_1^{\text{mult}}(\eta)\|^2] \leq \mathcal{M}_1 \|H_F^{1/2}\eta\| + 3\mathcal{M}_2 \|H_F^{1/2}\eta\|^2.$

Classical assumption

Hölder-type assumption (new)

$\mathcal{M}_1 = 0$ if the random field is linear,
$\mathcal{M}_1 \neq 0$ for quantization

Definition of $\mathfrak{C}$ania

Definition 2 (Additive and multiplicative noise)

Under the setting of (LSA), for any k in $\mathbb{N}^*$:

$$\xi_k^{\mathrm{add}} := \xi_k(0) \quad \textit{and} \quad \xi_k^{\mathrm{mult}} : z \in \mathbb{R}^d \mapsto \xi_k(z) - \xi_k^{\mathrm{add}}.$$

Assumption (Second moment of the multiplicative noise)

$\exists \mathcal{M}_1, \mathcal{M}_2 > 0$ *s.t. for any η in $\mathbb{R}^d$:*

1. $\mathbb{E}[\|\xi_1^{\mathrm{mult}}(\eta)\|^2] \leq 2\mathcal{M}_2 \|H_F^{1/2}\eta\|^2 + 4\mathcal{A}.$
2. $\mathbb{E}[\|\xi_1^{\mathrm{mult}}(\eta)\|^2] \leq \mathcal{M}_1 \|H_F^{1/2}\eta\| + 3\mathcal{M}_2 \|H_F^{1/2}\eta\|^2.$

Classical assumption

Hölder-type assumption (new)

$\mathcal{M}_1 = 0$ if the random field is linear,
$\mathcal{M}_1 \neq 0$ for quantization because:
$\mathbb{E}[\|\mathcal{C}(z) - \mathcal{C}(z')\|^2] \leq 12\sqrt{d}\min(\|z\|, \|z'\|)\|z - z'\| + 3(\omega + 1)\|z - z'\|^2$

Definition 2 (Additive and multiplicative noise)

Under the setting of (LSA), for any k in $\mathbb{N}^*$:

$$\xi_k^{\text{add}} := \xi_k(0) \quad \text{and} \quad \xi_k^{\text{mult}} : z \in \mathbb{R}^d \mapsto \xi_k(z) - \xi_k^{\text{add}}.$$

Assumption (Second moment of the multiplicative noise)

$\exists \mathcal{M}_1, \mathcal{M}_2 > 0$ *s.t. for any η in $\mathbb{R}^d$:*

1. $\mathbb{E}[\|\xi_1^{\text{mult}}(\eta)\|^2] \leq 2\mathcal{M}_2 \|H_F^{1/2}\eta\|^2 + 4\mathcal{A}.$
2. $\mathbb{E}[\|\xi_1^{\text{mult}}(\eta)\|^2] \leq \mathcal{M}_1 \|H_F^{1/2}\eta\| + 3\mathcal{M}_2 \|H_F^{1/2}\eta\|^2.$

Definition 3 (Ania's covariance.)

Under (LSA), we define the covariance of the additive noise: $\mathfrak{C}_{\text{ania}} = \mathbb{E}[\xi_1^{\text{add}} \otimes \xi_1^{\text{add}}].$

Definition 2 (Additive and multiplicative noise)

Under the setting of (LSA), for any k in $\mathbb{N}^*$:

$$\xi_k^{\text{add}} := \xi_k(0) \quad \text{and} \quad \xi_k^{\text{mult}} : z \in \mathbb{R}^d \mapsto \xi_k(z) - \xi_k^{\text{add}}.$$

Assumption (Second moment of the multiplicative noise)

$\exists \mathcal{M}_1, \mathcal{M}_2 > 0$ *s.t. for any* η *in* $\mathbb{R}^d$*:*

1. $\mathbb{E}[\|\xi_1^{\text{mult}}(\eta)\|^2] \leq 2\mathcal{M}_2 \|H_F^{1/2}\eta\|^2 + 4\mathcal{A}.$
2. $\mathbb{E}[\|\xi_1^{\text{mult}}(\eta)\|^2] \leq \mathcal{M}_1 \|H_F^{1/2}\eta\| + 3\mathcal{M}_2 \|H_F^{1/2}\eta\|^2.$

Definition 3 (Ania's covariance.)

Under (LSA), we define the covariance of the additive noise: $\mathfrak{C}_{\text{ania}} = \mathbb{E}[\xi_1^{\text{add}} \otimes \xi_1^{\text{add}}].$

Theorem 5 (Asymptotic result, from [PJ92])

Under some assumptions. Consider a sequence $(w_k)_{k \in \mathbb{N}^*}$ *produced in the setting of (LSA) for a step-size* $(\gamma_K)_{K \in \mathbb{N}^*}$ *s.t.* $\gamma_K = 1/\sqrt{K}$*. Then we have:*

$$\sqrt{K}(\overline{w}_K - w_*) \xrightarrow[K \to +\infty]{\mathcal{L}} \mathcal{N}(0, H_F^{-1} \mathfrak{C}_{\text{ania}} H_F^{-1}).$$

Convergence theorem

Theorem 6 ("Non-asymptotic convergence rate")

Under some assumptions. Consider a sequence $(w_k)_{k\in\mathbb{N}^*}$ produced by the setting of (LSA), for a constant step-size γ verifying some assumptions. Then for any horizon K, we have

$$\mathbb{E}[F(\overline{w}_{K-1}) - F(w_*)] \leq \frac{1}{2K}\left(\min\left(\frac{\|H_F^{-1/2}\eta_0\|}{\gamma\sqrt{K}}, \frac{\|\eta_0\|}{\sqrt{\gamma}}\right) + \sqrt{\operatorname{Tr}(\mathfrak{C}_{\text{ania}} H_F^{-1})} + O\left(\mu^{-1/2}\gamma^{1/4}\right) \right)^2 .$$

Theorem 6 ("Non-asymptotic convergence rate")

Under some assumptions. Consider a sequence $(w_k)_{k\in\mathbb{N}^*}$ produced by the setting of (LSA), for a constant step-size γ verifying some assumptions. Then for any horizon K, we have

$$\mathbb{E}[F(\overline{w}_{K-1}) - F(w_*)] \le \frac{1}{2K}\left(\min\left(\frac{\|H_F^{-1/2}\eta_0\|}{\gamma\sqrt{K}}, \frac{\|\eta_0\|}{\sqrt{\gamma}}\right) + \sqrt{\operatorname{Tr}(\mathfrak{C}_{\text{ania}} H_F^{-1})} + O(\mu^{-1/2}\gamma^{1/4}) \right)^2 .$$

Bias term, as in [BM13, DB15]

classical asymptotic noise term in CLT for (LSA)

asymptotically negligible for $\gamma = o(1)$, comes from multiplicative noise

$\eta_k = w_k - w_*$ | $\mathfrak{C}_{\text{ania}}$: additive noise's covariance | H_F: Hessian | $\mu = \min(\text{eig}(H_F))$

Theorem 6 ("Non-asymptotic convergence rate")

Under some assumptions. Consider a sequence $(w_k)_{k\in\mathbb{N}^*}$ produced by the setting of (LSA), for a constant step-size γ verifying some assumptions. Then for any horizon K, we have

$$\mathbb{E}[F(\overline{w}_{K-1}) - F(w_*)] \leq \frac{1}{2K}\left(\min\left(\frac{\|H_F^{-1/2}\eta_0\|}{\gamma\sqrt{K}}, \frac{\|\eta_0\|}{\sqrt{\gamma}} \right) + \sqrt{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}} H_F^{-1}\right) + O\left(\mu^{-1/2}\gamma^{1/4}\right)} \right)^2 .$$

Bias term, as in [BM13, DB15]

classical asymptotic noise term in CLT for (LSA)

asymptotically negligible for $\gamma = o(1)$, comes from multiplicative noise

Remarks:

- Asymptotically, the dominant term is $\sqrt{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}} H_F^{-1}\right)}$.
- Contrary to [BM13], the convergence rate *is not* necessarily independent of μ.
- Examining the explicit formulas of $\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}} H_F^{-1}\right)$ allows to determine the convergence rate.

$\eta_k = w_k - w_*$ | $\mathfrak{C}_{\mathrm{ania}}$: additive noise's covariance | H_F: Hessian | $\mu = \min(\mathrm{eig}(H_F))$

Computing $\mathrm{Tr}\left(\mathfrak{C}_{\text{ania}} H^{-1}\right)$

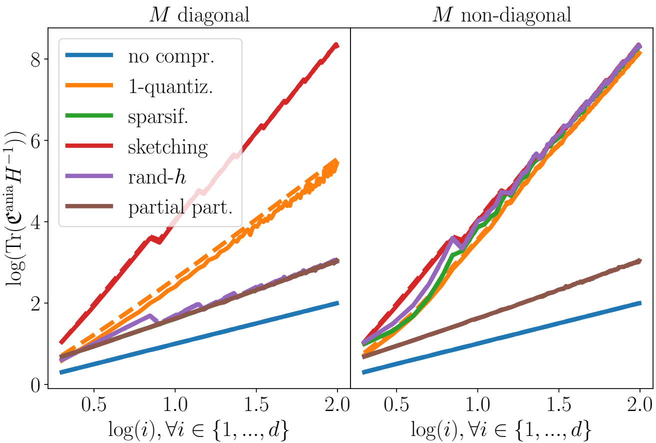
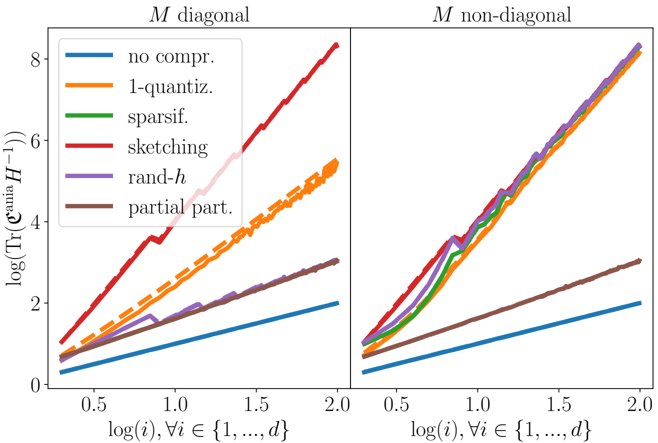


Figure 8: $\mathrm{Tr}\left(\mathfrak{C}_{\text{ania}} H^{-1}\right)$ - $K = 10^3, d \in [\![2, 100]\!]$, $D = \mathrm{Diag}\left((1/i^4)_{i=1}^d\right)$. Left: H diagonal. Right: H non-diagonal. (Plain line: empirical values; dashed lines: theoretical)

$\forall k \in \{1, \ldots, K\}, x_k \sim \mathcal{N}(0, H)$, with $H = QDQ^T$, $D = \mathrm{Diag}\left((1/i^4)_{i=1}^d\right)$ and Q an orthogonal matrix.



Depending on the compression scheme:

Classical LMS: $\mathfrak{C}_{\text{ania}} = H \quad (\times \sigma^2)$

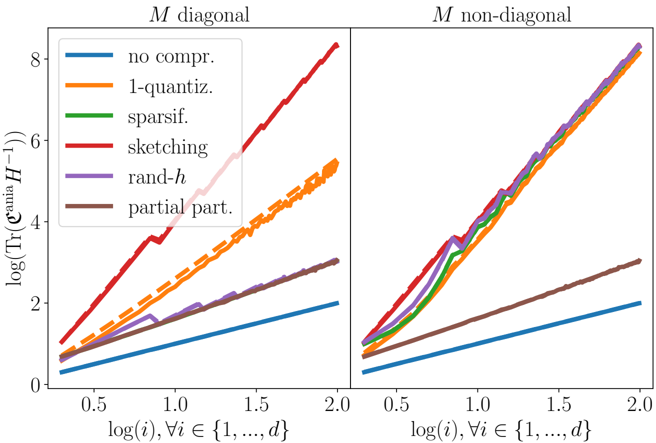
Partial part.: $\mathfrak{C}_{\text{ania}} = aH$

Sparsification: $\mathfrak{C}_{\text{ania}} = a' H + b\mathrm{Diag}(H)$

Sketching: $\mathfrak{C}_{\text{ania}} = a'' H + b' \mathrm{Tr}(H) \mathrm{I}_d$

Figure 8: $\mathrm{Tr}(\mathfrak{C}_{\text{ania}} H^{-1})$ - $K = 10^3, d \in [\![2, 100]\!]$, $D = \mathrm{Diag}\left((1/i^4)_{i=1}^d\right)$. Left: H diagonal. Right: H non-diagonal. (Plain line: empirical values; dashed lines: theoretical)

$\forall k \in \{1, \ldots, K\}, x_k \sim \mathcal{N}(0, H)$, with $H = QDQ^T$, $D = \mathrm{Diag}\left((1/i^4)_{i=1}^d\right)$ and Q an orthogonal matrix.



Depending on the compression scheme:

Classical LMS: $\mathfrak{C}_{\mathrm{ania}} = H \quad (\times \sigma^2)$

Partial part.: $\mathfrak{C}_{\mathrm{ania}} = aH$

Sparsification: $\mathfrak{C}_{\mathrm{ania}} = a' H + b\mathrm{Diag}(H)$

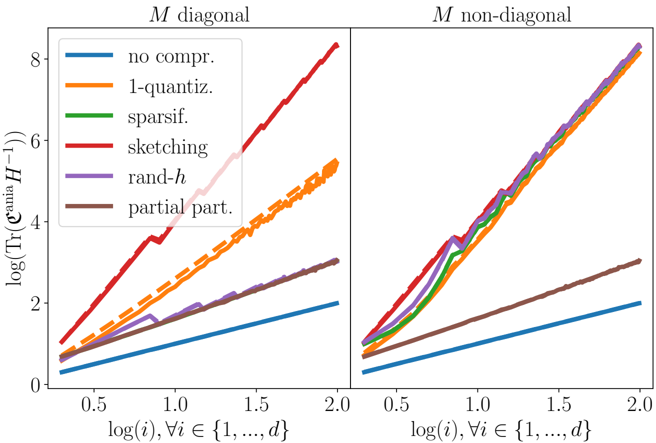
Sketching: $\mathfrak{C}_{\mathrm{ania}} = a'' H + b' \mathrm{Tr}(H) \mathrm{I}_d$

Structured noise

Isotropic noise

Figure 8: $\mathrm{Tr}(\mathfrak{C}_{\mathrm{ania}} H^{-1})$ - $K = 10^3, d \in [\![2, 100]\!]$, $D = \mathrm{Diag}\left((1/i^4)_{i=1}^d\right)$. Left: H diagonal. Right: H non-diagonal. (Plain line: empirical values; dashed lines: theoretical)

$\forall k \in \{1, \ldots, K\}, x_k \sim \mathcal{N}(0, H)$, with $H = QDQ^T$, $D = \mathrm{Diag}\left((1/i^4)_{i=1}^d\right)$ and Q an orthogonal matrix.



Depending on the compression scheme:

Classical LMS: $\mathfrak{C}_{\text{ania}} = H \qquad (\times \sigma^2)$

Partial part.: $\mathfrak{C}_{\text{ania}} = aH$

Sparsification: $\mathfrak{C}_{\text{ania}} = a'H + b\mathrm{Diag}(H)$

Sketching: $\mathfrak{C}_{\text{ania}} = a''H + b'\mathrm{Tr}(H)\mathrm{I}_d$

Structured noise

Isotropic noise

- Significantly impacts the limit distribution with a rate proportional to $\mathrm{Tr}(H^{-1})$.
- Same variance but different behaviors!

Figure 8: $\mathrm{Tr}(\mathfrak{C}_{\text{ania}} H^{-1})$ - $K = 10^3, d \in [\![2, 100]\!]$, $D = \mathrm{Diag}\left((1/i^4)_{i=1}^d\right)$. Left: H diagonal. Right: H non-diagonal. (Plain line: empirical values; dashed lines: theoretical)

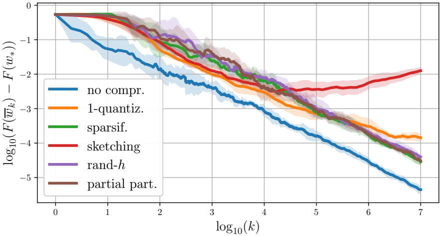
$\forall k \in \{1, \dots, K\}, x_k \sim \mathcal{N}(0, H)$, with $H = QDQ^T$, $D = \mathrm{Diag}\left((1/i^4)_{i=1}^d\right)$ and Q an orthogonal matrix.

Back to the comparison between various compressors in different scenarios

5 compressors: 4 scenarios, 4 different behaviors.

Back to the comparison between various compressors in different scenarios

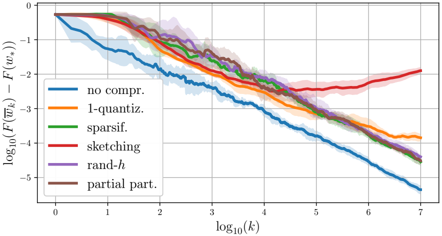
5 compressors: 4 scenarios, 4 different behaviors.



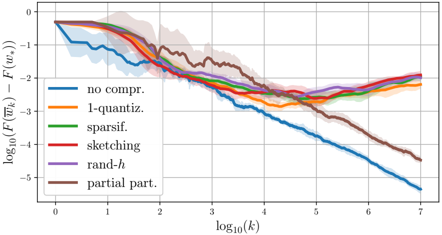
Fast eigenvalues' decay, diagonal covariance H.

Back to the comparison between various compressors in different scenarios

5 compressors: 4 scenarios, 4 different behaviors.



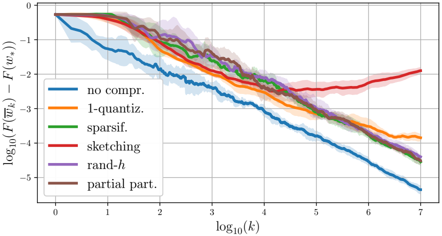
Fast eigenvalues' decay, diagonal covariance H.



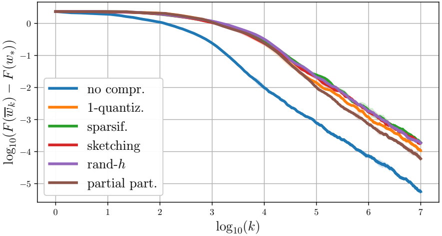
Fast eigenvalues' decay, non-diagonal covariance H.

Back to the comparison between various compressors in different scenarios

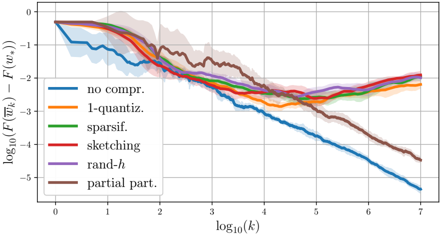
5 compressors: 4 scenarios, 4 different behaviors.



Fast eigenvalues' decay, diagonal covariance H.

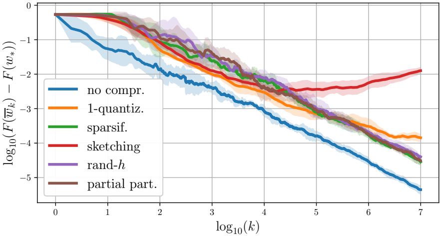


Slow eigenvalues' decay, non-diagonal covariance H.

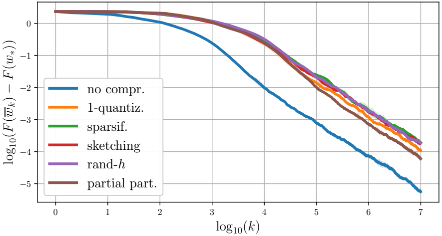


Fast eigenvalues' decay, non-diagonal covariance H.

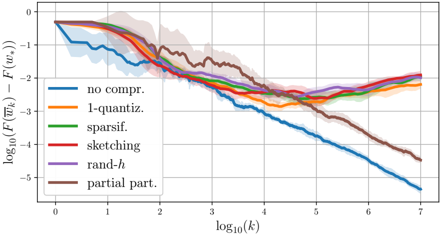
5 compressors: 4 scenarios, 4 different behaviors.



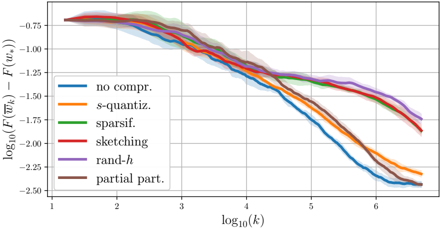
Fast eigenvalues' decay, diagonal covariance H.



Slow eigenvalues' decay, non-diagonal covariance H.



Fast eigenvalues' decay, non-diagonal covariance H.



Cifar10 with standardization (constant diagonal covariance H).

Partial conclusion

Summary of the contributions of the article:

- Analyze (LSA) under weak regularity assumptions of the noise field $(\xi_k)_k$.
- Provide a non-asymptotic theorem.
- Underline the key impact on convergence of the ania's covariance $\mathfrak{C}_{\text{ania}}$.
- Describe the link between, the compressor $\mathcal{C}$, the features' covariance H and the ania's covariance $\mathfrak{C}_{\text{ania}}$.
- Show how to compute the ania's covariance $\mathfrak{C}_{\text{ania}}$.
- Study the FL setting with heterogeneous clients.

Partial conclusion

Summary of the contributions of the article:

- Analyze (LSA) under weak regularity assumptions of the noise field $(\xi_k)_k$.
- Provide a non-asymptotic theorem.
- Underline the key impact on convergence of the ania's covariance $\mathfrak{C}_{\text{ania}}$.
- Describe the link between, the compressor $\mathcal{C}$, the features' covariance H and the ania's covariance $\mathfrak{C}_{\text{ania}}$.
- Show how to compute the ania's covariance $\mathfrak{C}_{\text{ania}}$.
- Study the FL setting with heterogeneous clients.

Examples of take-aways:

Take-away 6

- *Quantization not Lipschitz in squared expectation but satisfy a* ***Hölder-type*** *condition.*
- *Convergence degraded, yet achieve a* ***rate comparable to projection based compressors****.*

Take-away 7

- *Rand-1 and Partial Participation with probability* $(1/d)$*:* ***same variance condition****.*
- *But* ***PP is more robust*** *to ill conditioned problem.*

Conclusion

Conclusion of my thesis

Table 2: Summary of contributions.

	Bi-compr.	Heterogeneity	LSR	
I.	✓	✓		Interaction between compression and heterogeneity
II.	✓		(✓)	Asympt. cancels impact of down compression
III.		(✓)	✓	Beyond worst-case analysis

Conclusion of my thesis

Table 2: Summary of contributions.

	Bi-compr.	Heterogeneity	LSR	
I.	✓	✓		Interaction between compression and heterogeneity
II.	✓		(✓)	Asympt. cancels impact of down compression
III.		(✓)	✓	Beyond worst-case analysis

I. `Artemis` Bidirectional compression to reduce communication cost.
Key impact of memory on the convergence on non-i.i.d. data.

Conclusion of my thesis

Table 2: Summary of contributions.

	Bi-compr.	Heterogeneity	LSR	
I.	✓	✓		Interaction between compression and heterogeneity
II.	✓		(✓)	Asympt. cancels impact of down compression
III.		(✓)	✓	Beyond worst-case analysis

I. `Artemis` — Bidirectional compression to reduce communication cost.
Key impact of memory on the convergence on non-i.i.d. data.

II. `MCM` — Asympt, same rate of convergence as unidirectional compression.
Underlines the importance to not degrade the global model.

Conclusion of my thesis

Table 2: Summary of contributions.

	Bi-compr.	Heterogeneity	LSR	
I.	✓	✓		Interaction between compression and heterogeneity
II.	✓		(✓)	Asympt. cancels impact of down compression
III.		(✓)	✓	Beyond worst-case analysis

I. `Artemis` — Bidirectional compression to reduce communication cost.
Key impact of memory on the convergence on non-i.i.d. data.

II. `MCM` — Asympt, same rate of convergence as unidirectional compression.
Underlines the importance to not degrade the global model.

III. — Beyond the worst-case analysis of compression.
Analyze of the compressors' covariance.
Differences between compressors that have the same variance.

Thank you for your attention.

References

[AJL+23] A. Affouard, A. Joly, J. Lombardo, J. Champ, H. Goeau, M. Chouet, H. Gresse, C. Botella, and P. Bonnet. Pl@ntnet automatically identified occurrences. v1.8. Pl@ntNet, `https://plantnet.org/`, 2023.

[Bac14] Francis Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *The Journal of Machine Learning Research*, 15(1):595–627, 2014.

[Blu54] Julius R Blum. Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics*, pages 737–744, 1954.

[BM13] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Neural Information Processing Systems (NIPS)*, pages –, United States, December 2013.

[BMP12] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012.

[CG23] Chat-GPT3. Conversation with chat-gpt3. `https://chat.openai.com`, 2023. Accessed: 2023-08-16.

[DB15] Alexandre Defossez and Francis Bach. Averaged Least-Mean-Squares: Bias-Variance Trade-offs and Optimal Sampling Distributions. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 205–213. PMLR, February 2015. ISSN: 1938-7228.

[DBW12] John C. Duchi, Peter L. Bartlett, and Martin J. Wainwright. Randomized Smoothing for Stochastic Optimization. *SIAM Journal on Optimization*, 22(2):674–701, January 2012. Publisher: Society for Industrial and Applied Mathematics.

[DDB20] Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. *Ann. Statist.*, 48(3):1348–1382, 06 2020.

[DE23] Dall-E. An old castle on a cloud in a miyazaki style. `https://labs.openai.com/`, 2023. Accessed: 2023-04-01.

[DFB17] Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1):3520–3570, 2017. Publisher: JMLR. org.

[GLQ+19] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General Analysis and Improved Rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR, May 2019. ISSN: 2640-3498.

[GP23] Sébastien Gadat and Fabien Panloup. Optimal non-asymptotic analysis of the Ruppert–Polyak averaging stochastic algorithm. *Stochastic Processes and their Applications*, 156:312–348, February 2023.

[KT03] Vijay R Konda and John N Tsitsiklis. Linear stochastic approximation driven by slowly varying markov chains. *Systems & control letters*, 50(2):95–102, 2003.

[Lju77] Lennart Ljung. Analysis of recursive stochastic algorithms. *IEEE transactions on automatic control*, 22(4):551–575, 1977.

[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.

[LP21] Rémi Leluc and François Portier. SGD with Coordinate Sampling: Theory and Practice. *arXiv:2105.11818 [cs, stat]*, May 2021. arXiv: 2105.11818.

[LS83] Lennart Ljung and Torsten Söderström. *Theory and practice of recursive identification*. MIT press, 1983.

[MB11] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

[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.

[MMR+17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, April 2017. ISSN: 2640-3498.

[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.

[PD21] Constantin Philippenko and Aymeric Dieuleveut. Preserved central model for faster bidirectional compression in distributed settings. *Advances in Neural Information Processing Systems*, 34, 2021.

[PJ92] Boris Polyak and Anatoli Juditsky. Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization*, 30:838–855, July 1992.

[SBB+18] Kevin Scaman, Francis Bach, Sebastien Bubeck, Laurent Massoulié, and Yin Tat Lee. Optimal Algorithms for Non-Smooth Distributed Optimization in Networks. *Advances in Neural Information Processing Systems*, 31:2740–2749, 2018.

[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.

[Vil09] C. Villani. *Optimal transport : old and new*. Grundlehren der mathematischen wissenschaften. Springer, Berlin, 2009.

[ZHK19] Shuai Zheng, Ziyue Huang, and James Kwok. Communication-Efficient Distributed Blockwise Momentum SGD with Error-Feedback. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Open directions

- Evaluating the type and degree of heterogeneity within a network of clients.
- Compression and neural network: impact in a non-convex setting.
- New schemas of compression with independant coordinate compression.

Back-up on `Artemis`

Bulding statistical heterogeneous clients

Building non-i.i.d. *and* unbalanced datasets using a TSNE representation.

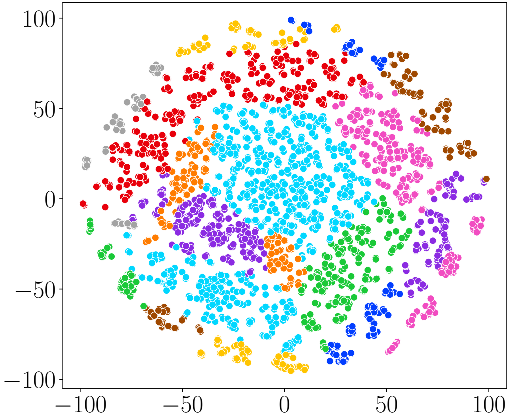


Figure 9: Superconduct

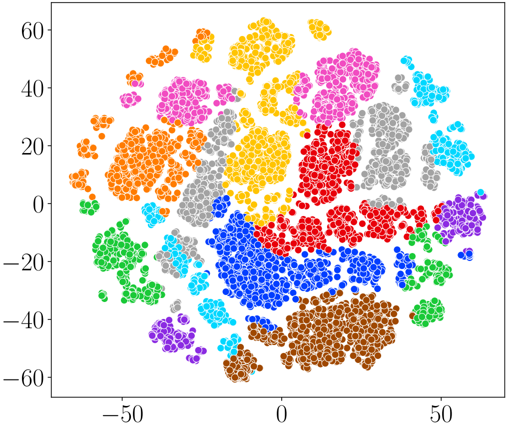


Figure 10: Quantum

A clue on the proof

We note $\tilde{g}_k = \mathcal{C}_{\mathrm{dwn}} \left(\frac{1}{N} \sum_{i=1}^N \mathcal{C}_{\mathrm{up}} (g_k^i - h_k^i) + h_k^i \right)$.

With no memory ($h_k^i = 0$ for any k in $\mathbb{N}^*$):

$$\mathbb{E} \left\| \tilde{g}_k \right\|^2 \leq \frac{A}{N^2} \sum_{i=0}^N \mathbb{E} \left\| g_k^i \right\|^2 + \frac{B}{N^2} \sum_{i=0}^N \mathbb{E} \left\| g_k^i - \nabla F_i(w_*) \right\|^2 + L \langle \nabla F(w_k), w_k - w_* \rangle \,.$$

With memory:

$$\mathbb{E} \left\| \tilde{g}_k \right\|^2 \leq \frac{A}{N^2} \sum_{i=1}^N \mathbb{E} \left\| g_k^i - g_{k,*}^i \right\|^2 + \frac{B}{N^2} \sum_{i=1}^N \mathbb{E} \left\| h_k^i - \nabla F_i(w_*) \right\|^2 + L \langle \nabla F(w_k), w_k - w_* \rangle + \frac{C \sigma_*}{N b} \,.$$

A clue on the proof

We note $\tilde{g}_k = \mathcal{C}_{\mathrm{dwn}}\left(\frac{1}{N}\sum_{i=1}^N \mathcal{C}_{\mathrm{up}}(g_k^i - h_k^i) + h_k^i\right)$.

With no memory ($h_k^i = 0$ for any k in $\mathbb{N}^*$):

$$\mathbb{E}\left\|\tilde{g}_k\right\|^2 \leq \frac{A}{N^2}\sum_{i=0}^N \mathbb{E}\left\|g_k^i\right\|^2 + \frac{B}{N^2}\sum_{i=0}^N \mathbb{E}\left\|g_k^i - \nabla F_i(w_*)\right\|^2 + L\langle \nabla F(w_k), w_k - w_*\rangle\,.$$

With memory:

$$\mathbb{E}\left\|\tilde{g}_k\right\|^2 \leq \frac{A}{N^2}\sum_{i=1}^N \mathbb{E}\left\|g_k^i - g_{k,*}^i\right\|^2 + \frac{B}{N^2}\sum_{i=1}^N \mathbb{E}\left\|h_k^i - \nabla F_i(w_*)\right\|^2 + L\langle \nabla F(w_k), w_k - w_*\rangle + \frac{C\sigma_*}{Nb}\,.$$

- $\langle \nabla F(w_k), w_k - w_*\rangle$ allows to use strong-convexity,
- $\left\|g_k^i\right\|^2$ makes appears the constant of heterogeneity B^2 !

Backup on MCM

A practical algorithm?

`Ghost` cannot be implemented in practice!

$\Longrightarrow$ Which choice do we have?

A practical algorithm?

Ghost cannot be implemented in practice!

$\Longrightarrow$ Which choice do we have?

Ghost

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

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

A practical algorithm?

Ghost cannot be implemented in practice!

$\Longrightarrow$ Which choice do we have?

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

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

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

A practical algorithm?

Ghost cannot be implemented in practice!

$\Longrightarrow$ Which choice do we have?

Ghost

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

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

Model compression

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

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

Update compression

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

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

A practical algorithm?

Ghost cannot be implemented in practice!

$\Longrightarrow$ Which choice do we have?

Ghost

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

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

Model compression

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

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

Update compression

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

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

Model difference compression

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

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

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

- Update compression
- Model difference compression
- Model compression
- MCM

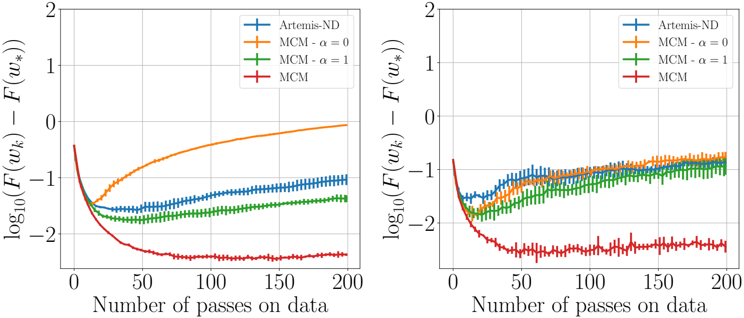


Figure 11: Comparing MCM on two datasets with three other algorithms using a non-degraded update, $\gamma = 1/L$.

Relation with randomized smoothing [DBW12, SBB+18]

Smoothed version of F:

$$F_\rho(w) :\mapsto \mathbb{E}[F(w + \rho X)], \text{ with } X \sim \mathcal{N}(0, I).$$

Relation with randomized smoothing [DBW12, SBB+18]

Smoothed version of F:

$$F_\rho(w) :\mapsto \mathbb{E}[F(w + \rho X)], \text{ with } X \sim \mathcal{N}(0, I).$$

$\nabla F(\hat{w}_{k-1})$ can be considered as an unbiased gradient of the smoothed function F_ρ at point w_{k-1}, with : $F_\rho : w \mapsto \mathbb{E}[F(w - w_{k-1} + \hat{w}_{k-1})]$

Relation with randomized smoothing [DBW12, SBB+18]

Smoothed version of F:

$$F_\rho(w) :\mapsto \mathbb{E}[F(w+\rho X)], \text{ with } X \sim \mathcal{N}(0, I).$$

$\nabla F(\hat{w}_{k-1})$ can be considered as an unbiased gradient of the smoothed function F_ρ at point w_{k-1}, with : $F_\rho : w \mapsto \mathbb{E}[F(w - w_{k-1} + \hat{w}_{k-1})]$ i.e.:

$$\nabla F(\hat{w}_{k-1}) = \nabla F_\rho(w_{k-1})$$

Relation with randomized smoothing [DBW12, SBB+18]

Smoothed version of F:

$$F_\rho(w) :\mapsto \mathbb{E}[F(w+\rho X)], \text{ with } X \sim \mathcal{N}(0,I).$$

$\nabla F(\hat{w}_{k-1})$ can be considered as an unbiased gradient of the smoothed function F_ρ at point w_{k-1}, with : $F_\rho : w \mapsto \mathbb{E}[F(w - w_{k-1} + \hat{w}_{k-1})]$ i.e.:

$$\nabla F(\hat{w}_{k-1}) = \nabla F_\rho(w_{k-1})$$

Then $\mathbb{E}\langle \nabla F(\hat{w}_{k-1}), w_{k-1} - w_* \rangle = \mathbb{E}\langle \nabla F_\rho(w_{k-1}), w_{k-1} - w_* \rangle$ which is the quantity that appears when developping the squared-norm of the update equation in the proof:

$$\mathbb{E}\|w_k - w_*\|^2 \leq \mathbb{E}\|w_{k-1} - w_*\|^2 - 2\gamma \langle \nabla F(\hat{w}_{k-1}), w_{k-1} - w_* \rangle + \gamma^2 \mathbb{E}\|\tilde{g}_k\|^2 .$$

Relation with randomized smoothing [DBW12, SBB+18]

Smoothed version of F:

$$F_\rho(w) :\mapsto \mathbb{E}[F(w+\rho X)], \text{ with } X \sim \mathcal{N}(0, I).$$

$\nabla F(\hat{w}_{k-1})$ can be considered as an unbiased gradient of the smoothed function F_ρ at point w_{k-1}, with : $F_\rho : w \mapsto \mathbb{E}[F(w - w_{k-1} + \hat{w}_{k-1})]$ i.e.:

$$\nabla F(\hat{w}_{k-1}) = \nabla F_\rho(w_{k-1})$$

Then $\mathbb{E}\langle \nabla F(\hat{w}_{k-1}), w_{k-1} - w_* \rangle = \mathbb{E}\langle \nabla F_\rho(w_{k-1}), w_{k-1} - w_* \rangle$ which is the quantity that appears when developping the squared-norm of the update equation in the proof:

$$\mathbb{E}\|w_k - w_*\|^2 \leq \mathbb{E}\|w_{k-1} - w_*\|^2 - 2\gamma \langle \nabla F(\hat{w}_{k-1}), w_{k-1} - w_* \rangle + \gamma^2 \mathbb{E}\|\tilde{g}_k\|^2 .$$

But two main differences:

- Objective function already smooth,
- Noise not Gaussian: we suffer from the noise because of compression and can not control it.

In more details ...

Let:

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

Theorem 7 (Convergence of MCM, convex case for any step-size γ)

Under all previous assumptions, for k in $\mathbb{N}^*$, for any $\gamma \leq \gamma_{\max}$, we have, for $\bar{w}_k = \frac{1}{k}\sum_{i=0}^{k-1} w_i$,

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

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

Comments on the variance term

For a constant γ,

- the variance term is upper bounded by

$$\frac{\gamma^2\sigma^2}{Nb}(\omega_{\mathrm{up}}+1)(1+64\gamma L{\omega_{\mathrm{dwn}}}^2)\,.$$

- impact of the downlink compression is attenuated by a factor γ. As γ decreases, this makes the limit variance similar to the one of Diana [MGTR19], i.e. without downlink compression:

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

- This is much lower than the variance for previous algorithms using double compression:

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

Comments on maximal step-size $\gamma_{\max}$

Maximal learning rate to ensure convergence:

$$\gamma_{\max} := \min(\gamma_{\max}^{\mathrm{up}}, \gamma_{\max}^{\mathrm{dwn}}, \gamma_{\max}^{\Upsilon})$$

where:

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

Comments on maximal step-size $\gamma_{\max}$

Maximal learning rate to ensure convergence:

$$\gamma_{\max} := \min(\gamma_{\max}^{\mathrm{up}}, \gamma_{\max}^{\mathrm{dwn}}, \gamma_{\max}^{\Upsilon})$$

where:

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

Remarks:

- constraints are weaker than in the "degraded" framework

$$\gamma_{\max}^{\mathsf{Dore}} \leq \left(8L(1+\omega_{\mathrm{dwn}})(1+\omega_{\mathrm{up}}/N)\right)^{-1},$$

- 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 $(L\omega^2)^{-1}$ for `Dore`/`Artemis`.
  Our $\gamma_{\max}$ is thus larger by a factor $\sqrt{\omega}$

Summary of rates and complexities

Rates, complexities, and maximal step size for `Diana`, `Artemis`, `Dore` and `MCM`.

Table 3: Summary of rates on the initial condition, limit variance, asympt. complexities and $\gamma_{\max}$.

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

$\Longrightarrow$ Consists in performing independent compressions for each device.

Theorem 8

Theorem 4 is still valid for Rand-MCM

- Improvement in Rand-MCM: because we average gradients at several random points, reducing the impact of ω_{dwn}.
- Dominating term is independent of ω_{dwn}: we expect to reduce only the second-order term.

$\Longrightarrow$ Consists in performing independent compressions for each device.

Theorem 8

Theorem 4 is still valid for Rand-MCM

- Improvement in Rand-MCM: because we average gradients at several random points, reducing the impact of ω_{dwn}.
- Dominating term is independent of ω_{dwn}: we expect to reduce only the second-order term.

Theorem 9 (Convergence in the quadratic case)

Under A1, A3, A7, with $\mu = 0$, if the function is quadratic, after running $K > 0$ iterations, for any $\gamma \leq \gamma_{\max}$, 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 $\Phi^{\mathrm{Rd}}(\gamma) = (1 + \omega_{\mathrm{up}}) \left(1 + \frac{4\gamma^2 L^2 \omega_{\mathrm{dwn}}}{K} \left(\frac{1}{\mathbf{C}} + \frac{\omega_{\mathrm{up}}}{N}\right)\right)$ *and* $\mathbf{C} = N$ *for Rand-MCM,* $\mathbf{C} = 1$ *for MCM.*

$\Longrightarrow$ Consists in performing independent compressions for each device.

Theorem 8

Theorem 4 is still valid for Rand-MCM

- Improvement in Rand-MCM: because we average gradients at several random points, reducing the impact of ω_{dwn}.
- Dominating term is independent of ω_{dwn}: we expect to reduce only the second-order term.

Theorem 9 (Convergence in the quadratic case)

Under A1, A3, A7, with $\mu = 0$, if the function is quadratic, after running $K > 0$ iterations, for any $\gamma \leq \gamma_{\max}$, 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 $\Phi^{\mathrm{Rd}}(\gamma) = (1 + \omega_{\mathrm{up}}) \left(1 + \frac{4\gamma^2 L^2 \omega_{\mathrm{dwn}}}{K} \left(\frac{1}{\mathbf{C}} + \frac{\omega_{\mathrm{up}}}{N}\right)\right)$ *and* $\mathbf{C} = N$ *for Rand-MCM,* $\mathbf{C} = 1$ *for MCM.*

- Quadratic functions: right hand term in Φ multiplied by an additional $\gamma\left(\frac{1}{\mathbf{C}} + \frac{\omega_{\mathrm{up}}}{N}\right)$.
- Randomization: further reduces by a factor N this term.

Backup on the compressors' covariance

Impact of the compression on the additive noise covariance

The additive noise writes for any $k \in \{1, \dots, K\}$, as:

$$\xi_k^{\text{add}} \overset{\text{def.}}{=}{}^2 \xi_k(0) \overset{\text{algo}}{=}{}^2 \nabla F(w_*) - \mathcal{C}_k(g_k(w_*)) = -\mathcal{C}_k((\langle x_k, w_* \rangle - y_k)x_k) = \mathcal{C}_k(\varepsilon_k x_k)\,.$$

By definition: $\mathfrak{C}_{\text{ania}} := \mathbb{E}[(\xi_k^{\text{add}})^{\otimes 2}] = \mathbb{E}[\mathcal{C}(\varepsilon_k x_k)^{\otimes 2}]$. Note also that $\mathcal{C}(\varepsilon_k x_k) \overset{\text{a.s.}}{=} \varepsilon_k \mathcal{C}(x_k)$ for all operators under consideration. Consequently

$$\mathfrak{C}_{\text{ania}} = \mathbb{E}[\varepsilon_k^2 \mathcal{C}(x_k)^{\otimes 2}] = \sigma^2 \mathbb{E}[\mathcal{C}(x_k)^{\otimes 2}]\,. \tag{3}$$

We study the covariance of $\mathcal{C}(x_k)$, for x_k a random variable with second-moment H, more generically we study the covariance of $\mathcal{C}(E)$, for E a random vector with distribution p_M with second moment $\mathbb{E}[E^{\otimes 2}] = M$.

Definition 4 (Compressor' covariance on p_M)

We define the following operator $\mathfrak{C}$ which returns the covariance of a random mechanism $\mathcal{C}$ acting on a distribution $p_M \in \mathcal{P}_M$,

$$\mathfrak{C}: \begin{array}{ccc} \mathbb{C} \times \mathcal{P}_M & \to & \mathbb{R}^{d \times d} \\ (\mathcal{C}, \quad p_M) & \to & \mathbb{E}[\mathcal{C}(E)^{\otimes 2}], \end{array}$$

where $E \sim p_M$ and the expectation is over the joint randomness of $\mathcal{C}$ and E, which are considered independent, that is $\mathbb{E}[\mathcal{C}(E)^{\otimes 2}] = \int_{\mathbb{R}^d} \mathbb{E}[\mathcal{C}(e)^{\otimes 2}] \mathrm{d}p_M(e)$.

Algorithm 3 (Distributed compressed LMS)

At any step k in $\{1,\dots,K\}$, each clients i in $\{1,\dots,N\}$ observes an oracle $g_k^i(\cdot)$ of the gradient of their local objective function F_i and applies a random compression mechanism $\mathcal{C}_k^i(\cdot)$.

For any step-size $\gamma > 0$ and any $k \in \mathbb{N}^*$, the resulting sequence of iterates $(w_k)_{k\in\mathbb{N}}$ satisfies:

$$w_k = w_{k-1} - \gamma \frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_k^i(g_k^i(w_{k-1})).$$

Equivalently, for $w \in \mathbb{R}^d$: $\xi_k(w) = \nabla F(w) - \frac{1}{N}\sum_{i=1}^{N} \mathcal{C}_k^i(g_k^i(w))$.

Application to Federated Learning

Algorithm 3 (Distributed compressed LMS)

At any step k in $\{1,\ldots,K\}$, each clients i in $\{1,\ldots,N\}$ observes an oracle $g_k^i(\cdot)$ of the gradient of their local objective function F_i and applies a random compression mechanism $\mathcal{C}_k^i(\cdot)$.

For any step-size $\gamma > 0$ and any $k \in \mathbb{N}^*$, the resulting sequence of iterates $(w_k)_{k\in\mathbb{N}}$ satisfies:

$$w_k = w_{k-1} - \gamma \frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_k^i(g_k^i(w_{k-1})).$$

Equivalently, for $w \in \mathbb{R}^d$: $\xi_k(w) = \nabla F(w) - \frac{1}{N}\sum_{i=1}^N \mathcal{C}_k^i(g_k^i(w))$.

Two scenarios:

- Heterogeneous covariances: for i, j in $\{1,\ldots,N\}$, possibly $H_i \neq H_j$ (covariate-shift),
- Heterogeneous optimal points: for i, j in $\{1,\ldots,N\}$, possibly $w_*^i \neq w_*^i$ (optimal-point-shift).

Algorithm 3 (Distributed compressed LMS)

At any step k in $\{1,\dots,K\}$, each clients i in $\{1,\dots,N\}$ observes an oracle $g_k^i(\cdot)$ of the gradient of their local objective function F_i and applies a random compression mechanism $\mathcal{C}_k^i(\cdot)$.

For any step-size $\gamma>0$ and any $k\in\mathbb{N}^*$, the resulting sequence of iterates $(w_k)_{k\in\mathbb{N}}$ satisfies:

$$w_k = w_{k-1} - \gamma \frac{1}{N}\sum_{i=1}^{N} \mathcal{C}_k^i(g_k^i(w_{k-1})).$$

Equivalently, for $w\in\mathbb{R}^d$: $\xi_k(w) = \nabla F(w) - \frac{1}{N}\sum_{i=1}^N \mathcal{C}_k^i(g_k^i(w))$.

Two scenarios:

- Heterogeneous covariances: for i,j in $\{1,\dots,N\}$, possibly $H_i \neq H_j$ (covariate-shift),
- Heterogeneous optimal points: for i,j in $\{1,\dots,N\}$, possibly $w_*^i \neq w_*^i$ (optimal-point-shift).

Corollary 1 (covariate-shift)

Theorem 6 holds.

How to compute the ania's covariance using the compressor's covariance?

We have for any clients $i, j \in \{1, \dots, N\}$, $w_*^i = w_*^j$, thus

$$\xi_k^{\text{add}} \overset{\text{def. 2}}{=} \xi_k(0) \overset{\text{algo 3}}{=} \nabla F(w_*) - \frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_k^i(g_k^i(w_*))$$

$$= -\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_k^i((\langle x_k^i, w_* \rangle - y_k^i) x_k^i) \underset{w_*^i = w_*^j}{=} \frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_k^i(\varepsilon_k^i x_k^i).$$

Next for all operators under consideration we have $\mathcal{C}_k^i(\varepsilon_k^i x_k^i) \overset{\text{a.s.}}{=} \varepsilon_k^i \mathcal{C}_k^i(x_k^i)$, thus, with p_{H_i} denoting the distribution of x_k^i with covariance H_i, we have:

$$\mathfrak{C}_{\text{ania}} = \mathbb{E}\left[(\xi_k^{\text{add}})^{\otimes 2}\right] = \mathbb{E}\left[\left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_k^i(\varepsilon_k^i x_k^i)\right)^{\otimes 2}\right] \overset{\text{indep. of } (\mathcal{C}_k^i)_{i=1}^d}{=} \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{E}\left[\mathcal{C}_k^i(\varepsilon_k^i x_k^i)^{\otimes 2}\right]$$

$$= \frac{\sigma^2}{N^2} \sum_{i=1}^{N} \mathbb{E}\left[\mathcal{C}_k^i(x_k^i)^{\otimes 2}\right] \overset{\text{Def. 4}}{=} \frac{\sigma^2}{N^2} \sum_{i=1}^{N} \mathfrak{C}(\mathcal{C}_k^i, p_{H_i}) \overset{\text{notation}}{=:} \frac{\sigma^2}{N} \overline{\mathfrak{C}((\mathcal{C}^i, p_{H_i})_{i=1}^N)}. \tag{4}$$

The operator $\overline{\mathfrak{C}((\mathcal{C}^i, p_{H_i})_{i=1}^N)}$ generalizes the notion of *compressor's covariance* (Definition 4).

Heterogeneous optimal points w_*^i 1/2

By definition, we have:

$$\xi_k(w-w_*) \overset{\text{Def. } 1\&\text{Alg.}3}{=} H_F(w-w_*) - \frac{1}{N}\sum_{i=1}^{N} \mathcal{C}^i(g_k^i(w)), \text{ thus } \xi_k^{\text{add}} \overset{\text{Def. } 2}{=} -\frac{1}{N}\sum_{i=1}^{N}\mathcal{C}^i(g_{k,*}^i),$$

with $g_{k,*}^i = (x_k^i \otimes x_k^i)(w_* - w_*^i) + x_k^i \varepsilon_k^i$. We thus have, for any $k \in \mathbb{N}$:

$$\begin{aligned}
\mathfrak{C}_{\text{ania}} &= \mathbb{E}\left[(\xi_k^{\text{add}})^{\otimes 2}\right] \overset{\nabla F(w_*)=0}{=} \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\mathcal{C}^i(g_{k,*}^i) - \nabla F_i(w_*)\right)^{\otimes 2}\right] \\
&\underset{\mathbb{E}\mathcal{C}_k^i(g_{k,*}^i)=\nabla F_i(w_*)}{\overset{\forall i \neq j,\ \mathcal{C}_k^i \perp \mathcal{C}_k^j}{=}} \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left(\mathcal{C}_k^i(g_{k,*}^i) - \nabla F_i(w_*)\right)^{\otimes 2}\right] \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\left(\mathbb{E}[\mathcal{C}_k^i(g_{k,*}^i)^{\otimes 2}] - \nabla F_i(w_*)^{\otimes 2}\right) \\
&= \frac{\sigma^2}{N^2}\sum_{i=1}^{N}\mathfrak{C}(\mathcal{C}^i, p_{\Theta_i}) - \frac{1}{N^2} H \sum_{i=1}^{N}(w_* - w_*^i)^{\otimes 2} H \preccurlyeq \frac{\sigma^2}{N}\overline{\mathfrak{C}((\mathcal{C}^i, p_{\Theta_i})_{i=1}^N)},
\end{aligned}$$

where p_{Θ_i} is the distribution of $g_{k,*}^i$ (for any k).

Heterogeneous optimal points w_*^i 2/2

In order to bound this quantity, following [DFB17], we make the following assumption.

Assumption 8

The kurtosis for the projection of the covariates x_1^i (or equivalently x_k^i for any k) is bounded on any direction $z \in \mathbb{R}^d$, i.e., there exists $\kappa > 0$, such that:

$$\forall i \in \{1, \dots, N\}, \ \forall z \in \mathbb{R}^d, \quad \mathbb{E}\left[\langle z, x_1^i \rangle^4\right] \leq \kappa \langle z, Hz \rangle^2$$

Proposition 1 (Impact of client-heterogeneity.)

Let W_* be a random variable uniformly distributed over $\{w_*^i, i \in \{1, \dots, N\}\}$, thus such that, $\mathrm{Cov}[W_*] = \frac{1}{N}\sum_{i=1}^N (w_* - w_*^i)^{\otimes 2}$, then:

$$\frac{1}{N}\sum_{i=1}^N \Theta_i \preccurlyeq \left(\kappa \mathrm{Tr}\left(H \mathrm{Cov}[W_*]\right) + \sigma^2\right) H.$$

1) Before compression is possibly applied, the noise remains structured, i.e., with covariance proportional to H, in the case of concept-shift

2) Compared to the homogeneous case, the averaged second-order moment increases from $\sigma^2 H$ to $(\kappa \mathrm{Tr}(H\mathrm{Cov}[W_*]) + \sigma^2) H$.
$\Longrightarrow$ shows impact of the dispersion of the optimal points. $(w_*^i)_{i=1}^N$.

Heterogeneous optimal points w_*^i with memory

Artemis with only uplink compression:

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

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

Heterogeneous optimal points w_*^i with memory

Artemis with only uplink compression:

$$
\begin{aligned}
w_k &= w_{k-1} - \gamma \frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_{\mathrm{up}}(g_k^i - h_k^i) + h_k^i \\
h_{k+1}^i &= h_k^i + \alpha \mathcal{C}_{\mathrm{up}}(g_k^i - h_k^i),
\end{aligned}
$$

⚠ Random fields are no more i.i.d. $\implies$ Definition 1 is no more fulfilled, invalidating Theorem 6. ⚠

Heterogeneous optimal points w_*^i with memory

Artemis with only uplink compression:

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

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

⚠ Random fields are no more i.i.d. $\Longrightarrow$ Definition 1 is no more fulfilled, invalidating Theorem 6. ⚠

Theorem 10 (CLT for concept-shift heterogeneity)

Under some assumption, with $\mu > 0$, for any step-size $(\gamma_k)_{k \in \mathbb{N}^*}$ s.t. $\gamma_k = 1/\sqrt{k}$. Then

1. $(\sqrt{K} \overline{\eta}_{K-1})_{K>0} \xrightarrow[K \to +\infty]{\mathcal{L}} \mathcal{N}(0, H_F^{-1} \mathfrak{C}_{\mathrm{ania}}^{\infty} H_F^{-1}),$
2. $\mathfrak{C}_{\mathrm{ania}}^{\infty} = \overline{\mathfrak{C}((\mathcal{C}^i, p_{\Theta_i'})_{i=1}^N)}$, *where* $p_{\Theta_i'}$ *is the distribution of* $g_{k,*}^i - \nabla F_i(w_*)$.

Heterogeneous optimal points w_*^i with memory

Artemis with only uplink compression:

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

⚠ Random fields are no more i.i.d. $\implies$ Definition 1 is no more fulfilled, invalidating Theorem 6. ⚠

Theorem 10 (CLT for concept-shift heterogeneity)

Under some assumption, with $\mu > 0$, for any step-size $(\gamma_k)_{k \in \mathbb{N}^*}$ s.t. $\gamma_k = 1/\sqrt{k}$. Then

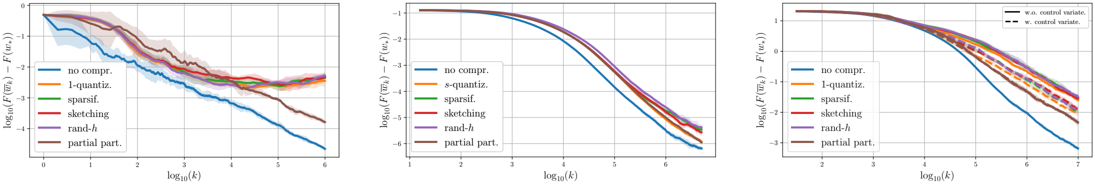
1. $(\sqrt{K}\eta_{K-1})_{K>0} \xrightarrow[K \to +\infty]{\mathcal{L}} \mathcal{N}(0, H_F^{-1} \mathfrak{C}_{\text{ania}}^{\infty} H_F^{-1})$,
2. $\mathfrak{C}_{\text{ania}}^{\infty} = \overline{\mathfrak{C}((\mathcal{C}^i, p_{\Theta_i'})_{i=1}^N)}$, *where $p_{\Theta_i'}$ is the distribution of $g_{k,*}^i - \nabla F_i(w_*)$.*

1. Settings of heterogeneous optimal points $(w_*^i)_{i=1}^N$: convergence still impacted by heterogeneity but with smaller additive noise's covariance as $\Theta_i' \prec \Theta_i$.
2. Deterministic gradients (batch case), we case $\Theta_i' \equiv 0$.
3. Recover asymptotically the results stated by Theorem 6 in the general setting of i.i.d. random fields $(\xi_k(\eta_{k-1}))_{k \in \mathbb{N}^*}$.

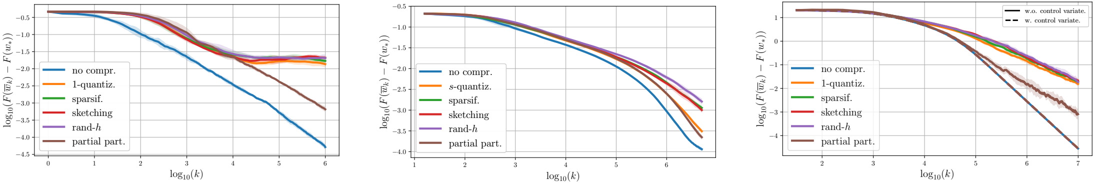
Experiments

Covariate-shift | Concept-shift

Synthetic dataset | Real datasets | Synthetic dataset



(a) No shift: $\forall i \in \{1,\dots,N\}, H_i = H$ **(b)** `quantum` **(c)** Batch stochastic gradient



(d) $\forall i, j \in \{1,\dots,N\}, H_i \neq H_j$ **(e)** `cifar-10` **(f)** True gradient $g_k^i = \nabla F_i$

Figure 12: Logarithm excess loss of the Polyak-Ruppert iterate iterations for $N = 10$ clients.