Compressed and distributed least-squares regression: convergence rates with applications to Federated Learning

Constantin Philippenko,
Inria Paris, Argo Team.

Work done while at CMAP, École Polytechnique, Institut Polytechnique de Paris.

Joint work with Aymeric Dieuleveut.

10ème édition des Journées Statistiques du Sud, 19-21 juin 2024, Toulouse.

Contents

General introduction on federated learning

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

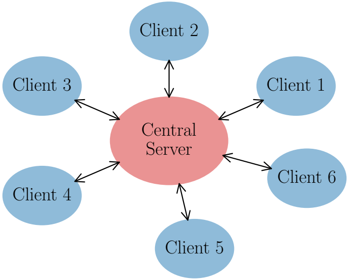
Figure 1: Automatic plant identification from photos using the mobile app [Pl@ntNet].

Paradigm: data is not centralized on a single location.

From a simple example to the challenges of Federated Learning

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

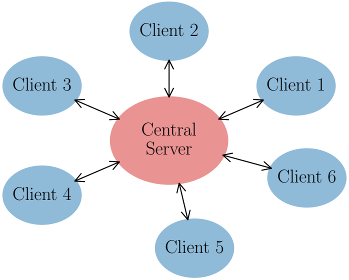
Paradigm: data is not centralized on a single location.



From a simple example to the challenges of Federated Learning

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

Paradigm: data is not centralized on a single location.



Privacy

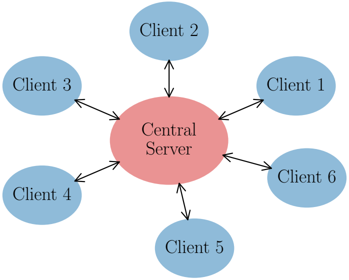
Communication cost

Data heterogeneity

From a simple example to the challenges of Federated Learning

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

Paradigm: data is not centralized on a single location.



Privacy

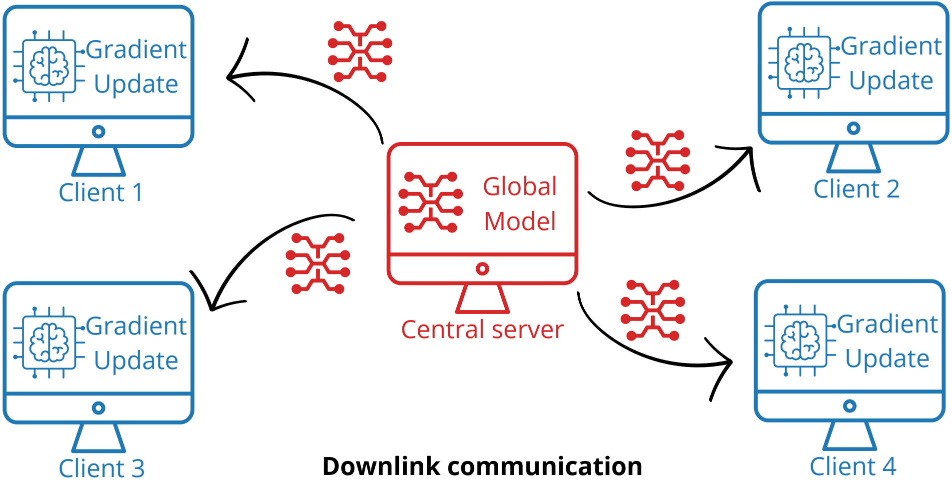
Communication cost

Data heterogeneity

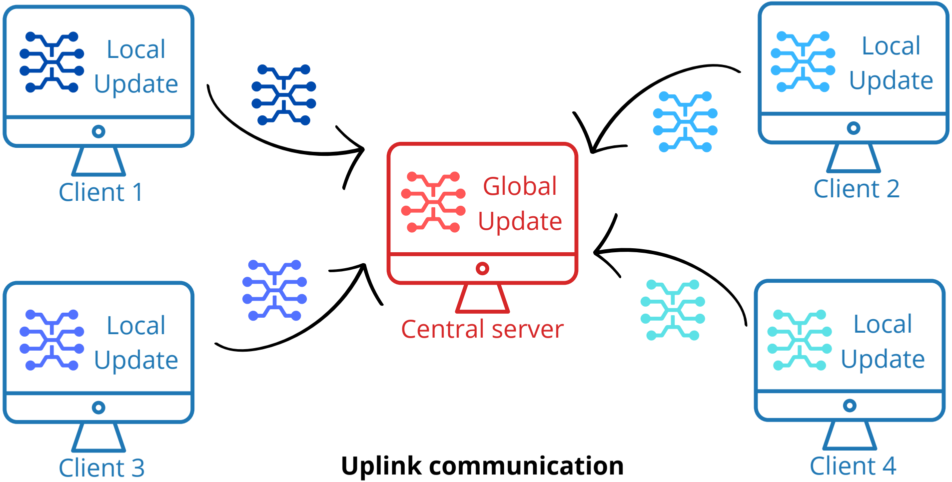
Goal of this presentation:

Focus simultaneously on two challenges: **reducing the cost of communication**.

Two steps of communication in FL



Downlink communication



Uplink communication

Applications of FL

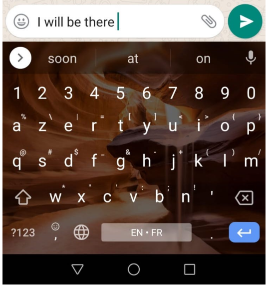


Figure 2: Gboards

- Google smart keyboard

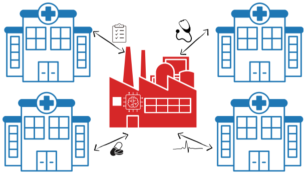


Figure 3: Hospitals collaboration

- Owkin (Substra)
- Inria and Université Côte d'Azur (FedBioMed)

Figure 4: Prediction of vehicle battery lifetime, pictures from AVL

- AVL (Research and Development for automotive industry)

Mathematical framework for compression

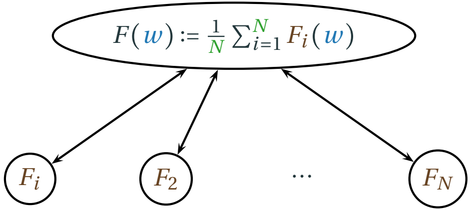
Goal : learning from a set of N clients [MMR+17]

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

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

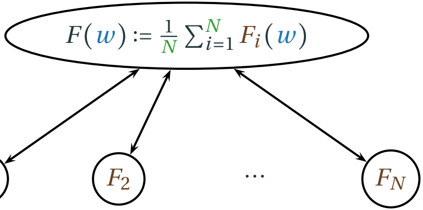
Goal : learning from a set of N clients [MMR+17]

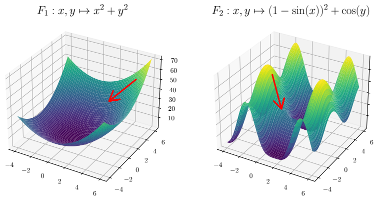
$$\min_{w \in \mathbb{R}^d} \left\{ F(w) := \frac{1}{N} \sum_{i=1}^{N} \underbrace{\mathbb{E}_{z \sim \mathcal{D}_i} \left[\ell(z, w) \right]}_{F_i(w)} \right\}.$$

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

Global loss

Local loss



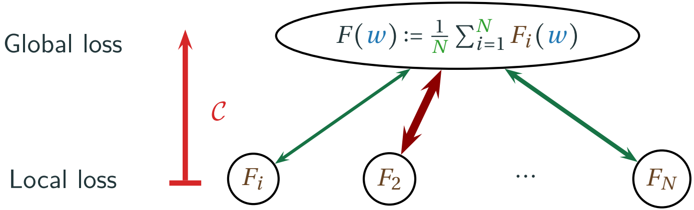


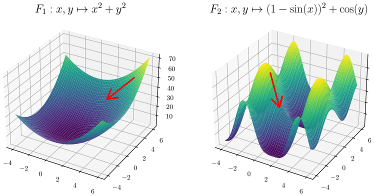
Distributed SGD: $\forall k \in \mathbb{N}, w_k = w_{k-1} - \gamma \left(\frac{1}{N} \sum_{i=1}^{N} g_k^i (w_{k-1}) \right).$

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





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

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

Exemples of compressors

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

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

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

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

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.

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

Big question: what is the impact of a compressor $\mathcal{C}$ 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 $\mathbb{R}^d$:

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

Big question: what is the impact of a compressor $\mathcal{C}$ 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.

Big question: what is the impact of a compressor $\mathcal{C}$ 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.

Big question: what is the impact of a compressor $\mathcal{C}$ 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:

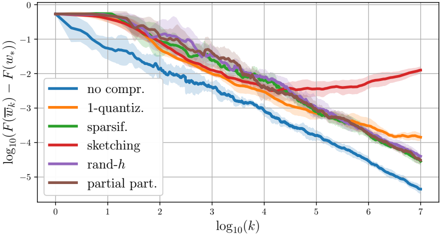
- $N = 1$ client.
- The client accesses K in $\mathbb{N}^*$ i.i.d. observations $(x_k, y_k)_{k \in \{1,\ldots,K\}} \sim \mathcal{D}^{\otimes K}$, such that there exists a well-defined model w_*:

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

5 compressors: 4 scenarios, 4 different behaviors.

Comparing various compressors in different scenarios

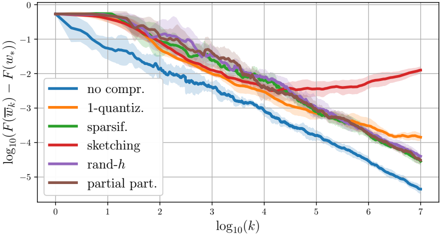
5 compressors: 4 scenarios, 4 different behaviors.



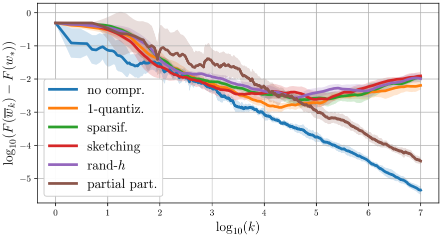
Sketching is very bad, quantiz. is slightly worse.

Comparing various compressors in different scenarios

5 compressors: 4 scenarios, 4 different behaviors.

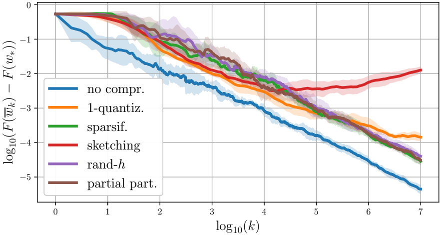


Sketching is very bad, quantiz. is slightly worse.

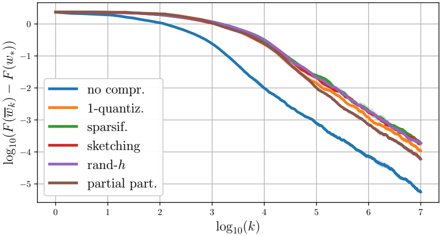


Only partial part. is good.

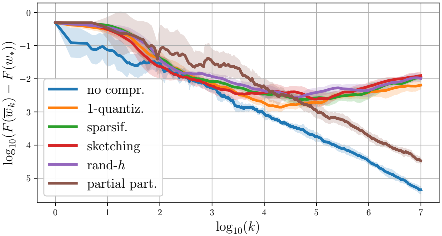
5 compressors: 4 scenarios, 4 different behaviors.



Sketching is very bad, quantiz. is slightly worse.

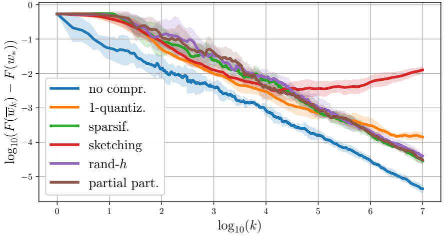


All compressors are equivalent and behave well.

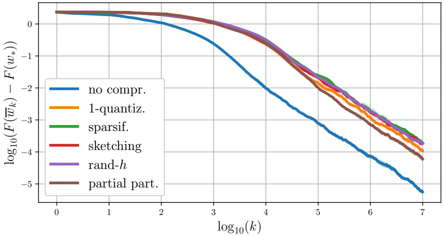


Only partial part. is good.

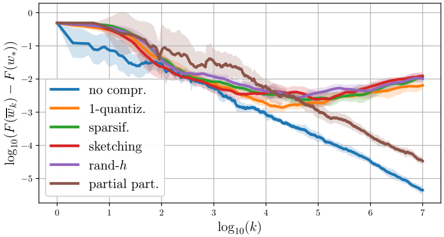
5 compressors: 4 scenarios, 4 different behaviors.



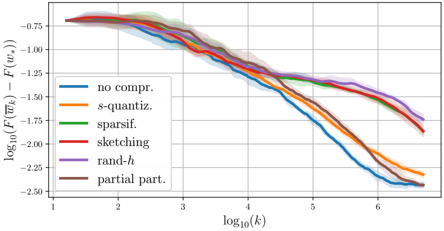
Sketching is very bad, quantiz. is slightly worse.



All compressors are equivalent and behave well.



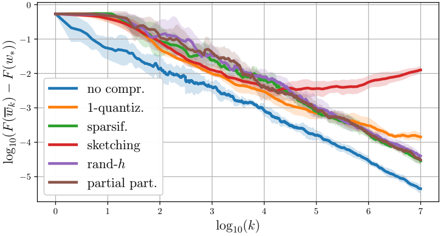
Only partial part. is good.

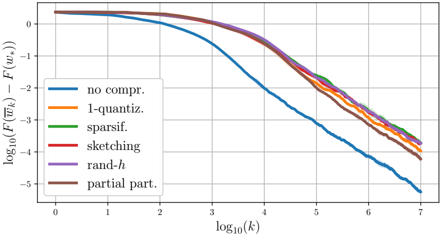


Quantiz. and partial part. are good.

5 compressors: 4 scenarios, 4 different behaviors.

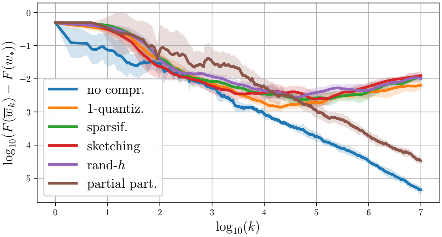
Can we explain this four different behaviors?

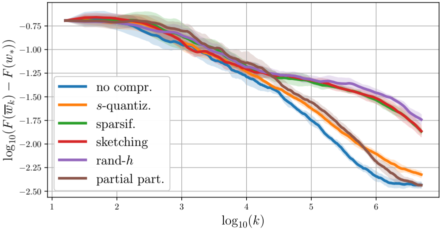




Sketching is very bad, quantiz. is slightly worse.

All compressors are equivalent and behave well.





Only partial part. is good.

Quantiz. and partial part. are good.

I. Non-asymptotic convergence result for (LSA)

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

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

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 $\{1, \ldots, K\}$, 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 $w \in \mathbb{R}^d$:

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

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 $\{1, \ldots, K\}$, 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 $w \in \mathbb{R}^d$:

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

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 $z, z' \in \mathbb{R}^d$

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

$\Longrightarrow$ 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 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 $\mathbb{N}^*$:

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

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

Classical assumption

Hölder-type assumption (new)

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

Definition 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)

$\mathcal{M}_1 = 0$ 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 $\mathbb{N}^*$:

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

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

Definition 3 (Ania's covariance.)

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

Theorem 1 (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:

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

Theorem 2 ("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 2 ("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

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

Bias term, as in [BM13, DB15]

classical asymptotic noise term in CLT for (LSA)

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

$\eta_k = w_k - w_*$

$\mathfrak{C}_{\text{ania}}$: additive noise's covariance

H_F: Hessian

$\mu = \min(\operatorname{eig}(H_F))$

Theorem 2 ("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

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

Bias term, as in [BM13, DB15]

classical asymptotic noise term in CLT for (LSA)

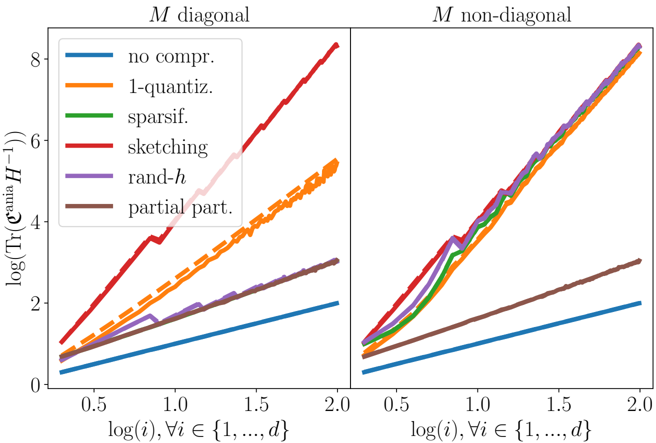
asymptotically negligible for $\gamma = o(1)$, comes from multiplicative noise

Remarks:

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

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

II. Compressed LSR on a single client



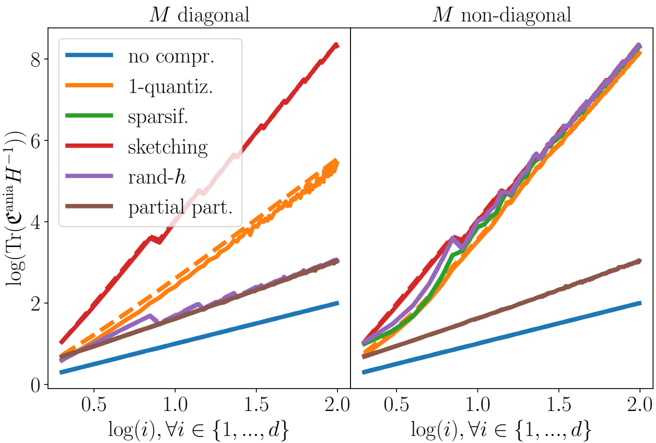
Depending on the compression scheme:

Classical LMS:	$\mathfrak{C}_{\text{ania}} = H$	$(\times\sigma^2)$
Partial part.:	$\mathfrak{C}_{\text{ania}} = aH$	
Sparsification:	$\mathfrak{C}_{\text{ania}} = a'H + b\mathrm{Diag}(H)$	
Rand-h:	$\mathfrak{C}_{\text{ania}} = b'\mathrm{Diag}(H)$	
Sketching:	$\mathfrak{C}_{\text{ania}} = a''H + b''\mathrm{Tr}(H)\mathrm{I}_d$	

Figure 5: $\mathrm{Tr}(\mathfrak{C}_{\text{ania}} H^{-1})$ - $K = 10^3, d \in [\![2, 100]\!]$, $D = \mathrm{Diag}\left((1/i^4)_{i=1}^d\right)$. 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.

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



Depending on the compression scheme:

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

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

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

Rand-h: $\mathfrak{C}_{\text{ania}} = b' \mathrm{Diag}(H)$

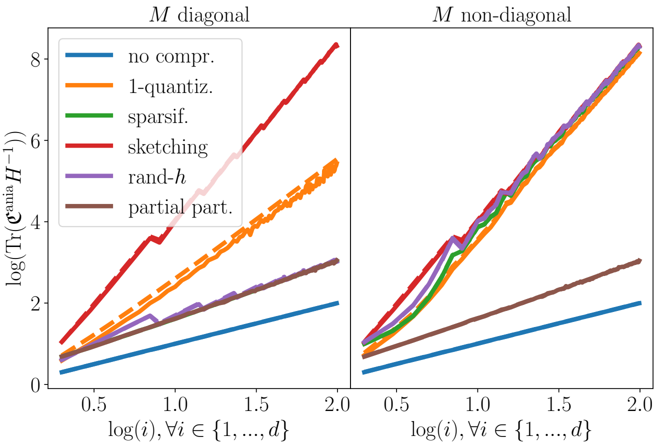
Sketching: $\mathfrak{C}_{\text{ania}} = a'' H + b'' \mathrm{Tr}(H) \mathrm{I}_d$

Structured noise

Isotropic noise

Figure 5: $\mathrm{Tr}(\mathfrak{C}_{\text{ania}} H^{-1})$ - $K = 10^3, d \in [\![2, 100]\!]$, $D = \mathrm{Diag}\left((1/i^4)_{i=1}^d\right)$. 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.



Depending on the compression scheme:

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

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

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

Rand-h: $\mathfrak{C}_{\mathrm{ania}} = b'\mathrm{Diag}(H)$

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

Structured noise

Isotropic noise

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

Figure 5: $\mathrm{Tr}(\mathfrak{C}_{\mathrm{ania}} H^{-1})$ - $K = 10^3, d \in [\![2, 100]\!]$, $D = \mathrm{Diag}\left((1/i^4)_{i=1}^d\right)$. Left: H diagonal. Right: H non-diagonal. (Plain line: empirical values; dashed lines: theoretical)

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

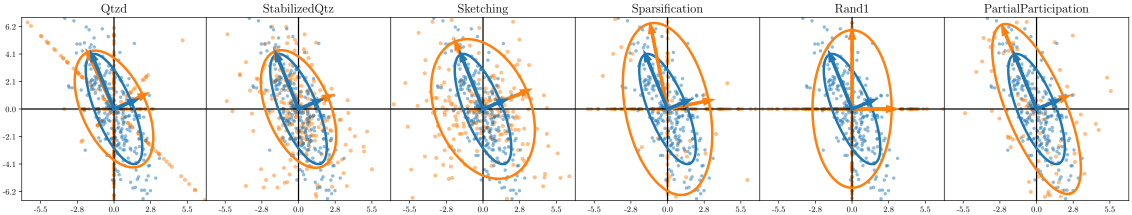


Figure 6: H not diagonal. Scatter plot of $(x_k)_{i=1}^K$/ $(\mathcal{C}(x_k))_{i=1}^K$ with its ellipse $\mathcal{E}_{\mathrm{Cov}[x_k]}/\mathcal{E}_{\mathrm{Cov}[\mathcal{C}(x_k)]}$.

$\forall k \in \{1, \ldots, K\}, x_k \sim \mathcal{N}(0, H)$, with $H = QDQ$, $D = \mathrm{Diag}(1, 10)$ and Q rotation matrix with angle $\pi/8$ in Figure 6.

Examples of take-aways

Take-away 1

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

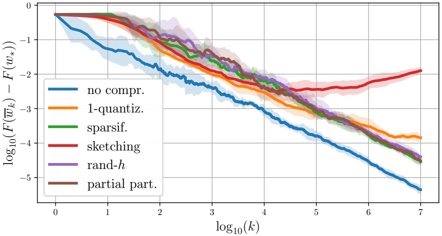
Take-away 2

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

5 compressors: 4 scenarios, 4 different behaviors.

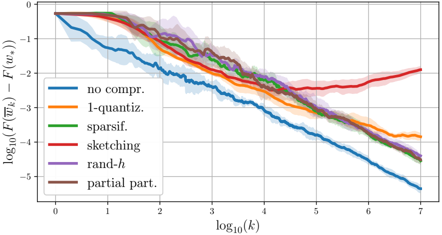
Back to the comparison between various compressors in different scenarios

5 compressors: 4 scenarios, 4 different behaviors.

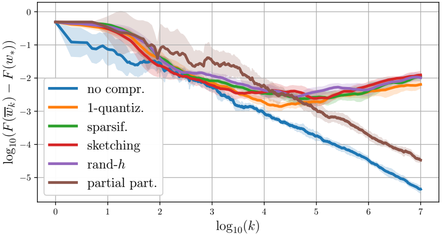


Fast eigenvalues' decay, diagonal covariance H.

5 compressors: 4 scenarios, 4 different behaviors.

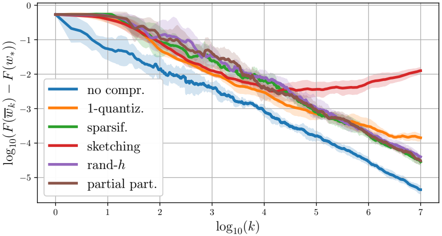


Fast eigenvalues' decay, diagonal covariance H.

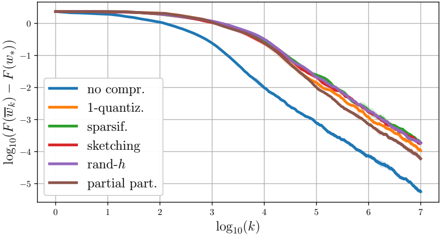


Fast eigenvalues' decay, non-diagonal covariance H.

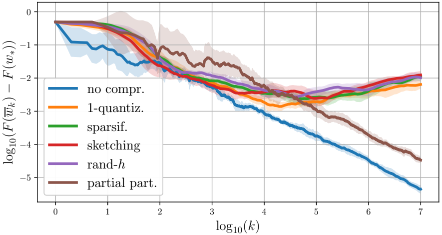
5 compressors: 4 scenarios, 4 different behaviors.



Fast eigenvalues' decay, diagonal covariance H.



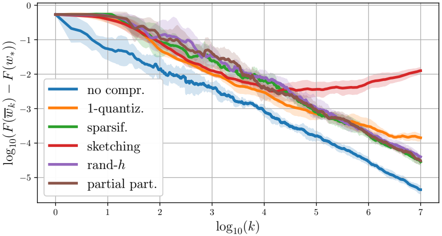
Slow eigenvalues' decay, non-diagonal covariance H.



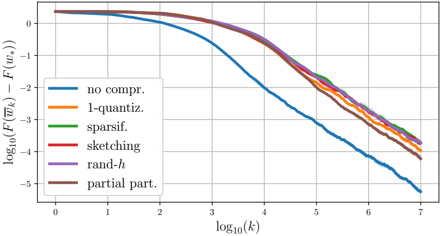
Fast eigenvalues' decay, non-diagonal covariance H.

Back to the comparison between various compressors in different scenarios

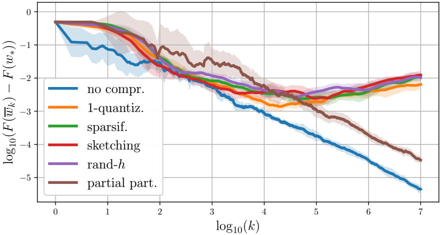
5 compressors: 4 scenarios, 4 different behaviors.



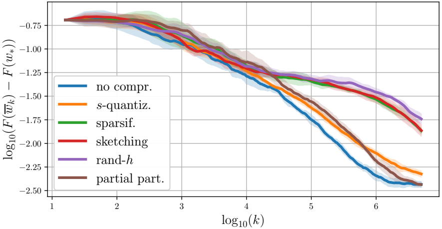
Fast eigenvalues' decay, diagonal covariance H.



Slow eigenvalues' decay, non-diagonal covariance H.



Fast eigenvalues' decay, non-diagonal covariance H.



Cifar10 with standardization (constant diagonal covariance H).

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 $\mathfrak{C}_{\mathrm{ania}}$.
- Describe the link between, the compressor $\mathcal{C}$, the features' covariance H and the ania's covariance $\mathfrak{C}_{\mathrm{ania}}$.
- Show how to compute the ania's covariance $\mathfrak{C}_{\mathrm{ania}}$.
- Study the FL setting with heterogeneous clients.

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.

Take-away 3

- *Beyond the worst-case analysis of compression.*
- *Analyze of the compressors' covariance $\mathfrak{C}$.*
- *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.

[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.

[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.

[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.

[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.

[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.

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