Journal of Machine Learning Research 25 (2024) 1-80 Submitted 08/23; Published 09/24

Compressed and distributed least-squares regression: convergence rates with applications to federated learning

Constantin Philippenko
CONSTANTIN.PHILIPPENKO@POLYTECHNIQUE.EDU
École polytechnique, Institut Polytechnique de Paris, CMAP

Aymeric Dieuleveut
AYMERIC.DIEULEVEUT@POLYTECHNIQUE.EDU
École polytechnique, Institut Polytechnique de Paris, CMAP

Editor: Lorenzo Rosasco

Abstract

In this paper, we investigate the impact of compression on stochastic gradient algorithms for machine learning, a technique widely used in distributed and federated learning. We underline differences in terms of convergence rates between several unbiased compression operators, that all satisfy the same condition on their variance, thus going beyond the classical worst-case analysis. To do so, we focus on the case of least-squares regression (LSR) and analyze a general stochastic approximation algorithm for minimizing quadratic functions relying on a random field. We consider weak assumptions on the random field, tailored to the analysis (specifically, expected Hölder regularity), and on the noise covariance, enabling the analysis of various randomizing mechanisms, including compression. We then extend our results to the case of federated learning.

More formally, we highlight the impact on the convergence of the covariance $\mathfrak{C}_{\mathrm{ania}}$ of the <u>a</u>*dditive* <u>n</u>*oise* <u>i</u>*nduced by the* <u>a</u>*lgorithm*. We demonstrate despite the non-regularity of the stochastic field, that the limit variance term scales with $\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}} H_F^{-1}\right)/K$ (where H_F is the Hessian of the optimization problem and K the number of iterations) generalizing the rate for the vanilla LSR case where it is $\sigma^2 \mathrm{Tr}\left(H_F H_F^{-1}\right)/K = \sigma^2 d/K$ (Bach and Moulines, 2013). Then, we analyze the dependency of $\mathfrak{C}_{\mathrm{ania}}$ on the compression strategy and ultimately its impact on convergence, first in the centralized case, then in two heterogeneous FL frameworks.

Keywords: Large-scale optimization, linear stochastic approximation, least-squares regression, federated learning, compression

1. Introduction

Large-scale optimization (Bottou and Bousquet, 2007) has become ubiquitous in today's learning problems due to the incredible growth of data collection. It becomes computationally extremely hard to process a full dataset or even, to store it on a single device (Abadi et al., 2016; Seide and Agarwal, 2016; Caldas et al., 2019). This led practitioners to either process each observation only once in a streaming fashion, or to design distributed algorithms. This paper is part of this line of work and considers in particular stochastic federated algorithms (Konečný et al., 2016; McMahan et al., 2017) that use a central server to orchestrate the training over a network of N in $\mathbb{N}^*$ clients.

A well-identified challenge in this framework is the communication cost of the learning process (Seide et al., 2014; Chilimbi et al., 2014; Strom, 2015) based on stochastic gradient algorithms. Indeed, iteratively exchanging gradient or model information between the local workers and the central server generates a huge computational and bandwidth bottleneck. To reduce this communication cost, two strategies have been widely implemented and analyzed: performing local updates (see e.g. McMahan et al., 2017; Karimireddy et al., 2020), or reducing the size of the exchanged messages by passing them through a compression operator, on the uplink channel (Seide et al., 2014; Alistarh et al., 2017, 2018; Mishchenko et al., 2019; Karimireddy et al., 2019; Wu et al., 2018; Horvath et al., 2022; Mishchenko et al., 2019; Li et al., 2020; Richtarik et al., 2021), or on both uplink and downlink channels (Harrane et al., 2018; Tang et al., 2019; Liu et al., 2020; Zheng et al., 2019; Philippenko and Dieuleveut, 2020, 2021; Gorbunov et al., 2020b; Sattler et al., 2019; Fatkhullin et al., 2021). These two strategies, although typically analyzed independently, are often combined. We focus on compression; to reduce the cost of exchanging a vector, three techniques are combined: (1) sending the message to only a few clients, (2) sending only a fraction of the coordinates, (3) sending low-precision updates.

Most analyses of the impact of compression schema rely on generic assumptions on the compression operator $\mathcal{C}$, typically either *contractive*, i.e. for any z in $\mathbb{R}^d$, $\|\mathcal{C}(z) - z\| < (1-\delta)\|z\|$ with $\delta \in]0;1[$ (almost surely or in expectation, see for instance Seide et al., 2014; Stich et al., 2018; Karimireddy et al., 2019; Ivkin et al., 2019; Koloskova et al., 2019; Gorbunov et al., 2020b; Beznosikov et al., 2020), or unbiased with bounded variance increase, i.e., for any z in $\mathbb{R}^d$, $\mathbb{E}[\mathcal{C}(z)] = z$ and $\mathbb{E}[\|\mathcal{C}(z) - z\|^2] \leq \omega \|z\|^2$ for a parameter $\omega > 1$ (see among others Alistarh et al., 2017; Wu et al., 2018; Mishchenko et al., 2019; Chraibi et al., 2019; Gorbunov et al., 2020a; Reisizadeh et al., 2020; Horvath et al., 2022; Kovalev et al., 2021; Philippenko and Dieuleveut, 2020, 2021; Haddadpour et al., 2021; Li and Richtárik, 2021; Khirirat et al., 2018). Unlike biased—and often deterministic—operators, unbiased operators typically benefit from a variance reduction proportional to the number of clients (e.g., Gorbunov et al., 2020b vs Horváth et al., 2019).

In parallel, a line of work has thus focused on the design of compression schemes satisfying one of these two assumptions (Bernstein et al., 2018; Dai et al., 2019; Beznosikov et al., 2020; Horvath et al., 2022; Xu et al., 2020; Leconte et al., 2021; Gandikota et al., 2021; Ramezani-Kebrya et al., 2021; Horvath et al., 2022). Two fundamental strategies are typically combined: (1) quantization (Rabbat and Nowak, 2005; Gersho and Gray, 2012; Alistarh et al., 2018), and (2) random projection (Vempala, 2005; Rahimi and Recht, 2008; Nesterov, 2012; Nutini et al., 2015). These methods are compared based on (1) the number of bits required for storing or exchanging a d dimensional vector and (2) the resulting variance increase ω or contractiveness constant δ. Consequently, convergence results are *worst-case* results over the class of compression operators: two compression operators satisfying the same variance assumption are regarded as producing the same convergence rate.

The goal of this paper is to provide an in-depth analysis of compression within a fundamental learning framework, namely least-squares regression (LSR, Legendre, 1806), in order to highlight the differences in convergence between several unbiased compression schemes having the *same* variance increase.

More precisely, we aim at analyzing updates on a sequence of models $(w_k)_{k\in\mathbb{N}}$ of the form $w_k = w_{k-1} - \frac{\gamma}{N}\sum_{i=1}^N \mathcal{C}^i(g_k^i(w_{k-1}))$, where γ is the step-size and g_k^i is a stochastic oracle on the gradient of the least-squares objective function of client i (see Algorithms 2 and 3).

To the best of our knowledge, this study is the first to *compare compressors that are in the same class*, i.e. satisfying the same variance assumption. Especially, this analysis will highlight the impact of (1) the compression scheme's regularity (Lipschitz in squared expectation or not) and of (2) the correlation between the compression of the different coordinates. We highlight three examples of possible take-aways from our analysis, that will be detailed in Section 3.

Take-away 1 *Quantization-based compression schemes do not have* Lipschitz in squared expectation *regularity but satisfy a Hölder condition. Because of that, their convergence is degraded, yet they asymptotically achieve a rate comparable to projection-based compressors, in which the limit covariance is similar.*

Take-away 2 *Rand-h and partial participation with probability* (h/d) *satisfy the same variance condition. Yet the convergence of compressed least mean squares algorithms for partial participation is more robust to ill-conditioned problems.*

Take-away 3 *The asymptotic convergence rate is expected to be at least as good for quantization than for sparsification or randomized coordinate selection,* if *the features are standardized. On the contrary, if the features are independent and the feature vector is normalized, then quantization is worse than sparsification or randomized coordinate selection.*

We consider a random-design LSR framework and make the following assumption on the input-output pairs distribution

Model 1 (Federated case) *We consider N clients. Each client i in $[N]$ accesses K in $\mathbb{N}^*$ i.i.d. observations $(x_k^i, y_k^i)_{k\in[K]} \sim \mathcal{D}_i^{\otimes K}$, such that there exists a well-defined client-dependent model w_*^i:*

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

for an i.i.d. sequence $\left((\varepsilon_k^i)_{k\in[K],i\in[N]}\right)$ independent of $\left((x_k^i)_{k\in[K],i\in[N]}\right)$. We use the generic notation $(x^i, y^i, \varepsilon^i)$ for such an input-output-noise triplet on client i. Moreover, we assume that the inputs' second moment[1] *is bounded to define $\mathbb{E}[x^i \otimes x^i] = H_i$ and $\mathbb{E}[\|x^i\|^2] = R_i^2$; such that $\mathbb{E}[\|x^i\|^2 x^i \otimes x^i] \preccurlyeq R_i^2 H_i$. For any $i \in [N]$, we consider the expected squared loss on client i of a model w as $F_i(w) := \frac{1}{2}\mathbb{E}_{(x^i,y^i)\sim\mathcal{D}_i}[(\left\langle x^i, w\right\rangle - y^i)^2]$.*

Remark 1 (Almost surely bounded features) *In the case of linear compressors, we will also assume that for each client i in $[N]$, features are almost surely bounded by R_i^2.*

This model is classical in the single worker case (e.g. Hsu et al., 2012; Bach and Moulines, 2013):

1. In the following, we may refer to this matrix H as the *covariance* (in the case of centered features, covariance is equal to the second moment)

Model 2 (Centralized case) *We consider Model 1 with $N = 1$ client. For simplicity, we then omit the i superscript.*

We focus on the problem of minimizing the global expected risk $F : \mathbb{R}^d \to \mathbb{R}$, thus finding the optimal model w_* in $\mathbb{R}^d$ such that:

$$w_* = \arg\min_{w \in \mathbb{R}^d} \left\{ F(w) := \frac{1}{N} \sum_{i=1}^{N} F_i(w) \right\} \tag{OPT}$$

Note that we assume that $\mathrm{Span}\{\mathrm{Supp}(x^i), i \in [N]\} = \mathbb{R}^d$ to ensure the existence and uniqueness of w_*.

The empirical version of the risk minimization admits an explicit formula, yet is computationally too expensive to compute for large problems. This is why, in practice, LSR is solved using iterative stochastic algorithms, for example Stochastic Gradient Descent (SGD, see Robbins and Monro, 1951). SGD for LSR is often referred to as the *Least Mean Squares* (LMS) algorithm (Bershad, 1986; Macchi, 1995). Analysis of LMS (Györfi and Walk, 1996; Bach and Moulines, 2013) and its variants received a lot of interest over the last decades. Indeed despite its simplicity, LSR is a model of choice for practitioners because of its efficiency to train good and interpretable models (see e.g. Molnar, 2018, chapter 5.1). Moreover, its simplicity enables to isolate and analyze challenges faced in specific configurations, for instance, non-strong convexity (Bach and Moulines, 2013), interaction between acceleration and stochasticity (Dieuleveut et al., 2017; Jain et al., 2018a; Varre and Flammarion, 2022), non-uniform iterate averaging (Jain et al., 2018b; Neu and Rosasco, 2018; Muecke et al., 2019), infinite-dimensional frameworks (Dieuleveut and Bach, 2016), or over-parametrized regimes and double descent phenomena (Belkin et al., 2019).

Our approach follows this line of work: our goal is to analyze the impact of *compression* in FL algorithms, by providing a careful study of compressed LMS, based on a fine-grained analysis of Stochastic Approximation (SA) under weak assumptions on the random field. More precisely, we consider linear stochastic approximation recursion, to find a zero of the linear mean-field ∇F.

Definition 2 (Linear Stochastic Approximation, LSA) *Let $w_0 \in \mathbb{R}^d$ be the initialization, the linear[2] stochastic approximation recursion is defined as:*

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

where $\gamma > 0$ is the step-size and $(\xi_k)_{k \in \mathbb{N}^*}$ is a sequence of i.i.d. zero-centered random fields that characterizes the stochastic oracle on $\nabla F(\cdot)$. For any $k \in \mathbb{N}^*$, we denote $\mathcal{F}_k = \sigma(\xi_1, \dots, \xi_k)$, such that the filtration $(\mathcal{F}_k)_{k \geq 0}$ is adapted to $(w_k)_{k \geq 0}$.

We assume that F is quadratic, we denote H_F its Hessian, $R_F^2 := \mathrm{Tr}(H_F)$ its trace and μ its smallest eigenvalue. For any k in $\mathbb{N}$, with $\eta_k = w_k - w_*$, we get equivalently:

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

2. 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.

As underlined by Bach and Moulines (2013), (LSA) corresponds to a homogeneous Markov chain. A study of stochastic approximation using results and techniques from the Markov chain literature can be found for instance in Freidlin and Wentzell (1998) or more recently in Dieuleveut et al. (2020).

(LSA) encompasses three examples of interest, the first one is the classical LMS algorithm. Indeed, with the observations in Models 1 and 2, for any client $i \in [N]$, any iteration k in $[K]$, any model $w \in \mathbb{R}^d$,

$$g_k^i(w) := (\langle x_k^i, w \rangle - y_k^i) x_k^i \tag{2}$$

is an unbiased oracle of $\nabla F_i(w)$. This can be used to define the following three algorithms.

Algorithm 1 (LMS) *For LMS algorithm, with a single worker (Model 2), we have for all $k \in \mathbb{N}$, $w_k = w_{k-1} - \gamma g_k(w_{k-1}) = w_{k-1} - \gamma(\langle x_k, w_{k-1} \rangle - y_k) x_k$, thus equivalently, we have $\xi_k(\cdot) = (\mathbb{E}[x_1 x_1^\top] - x_k x_k^\top)(\cdot) + \epsilon_k x_k$. Indeed, for any w in $\mathbb{R}^d$, $\xi_k(w - w_*) = \nabla F(w) - g_k(w) = \mathbb{E}[x_1 x_1^\top](w - w_*) - (\langle x_k, w \rangle - y_k) x_k = (\mathbb{E}[x_1 x_1^\top] - x_k x_k^\top)(w - w_*) - (\langle x_k, w_* \rangle - y_k) x_k$.*

Second, the case of a single client compressed LMS algorithm.

Algorithm 2 (Centralized compressed LMS) *A single client ($N = 1$) observes at any step $k \in [K]$ an oracle $g_k(\cdot)$ on the gradient of the objective function F, and applies a random compression mechanism $\mathcal{C}_k(\cdot)$. Thus, for any step-size $\gamma > 0$ and any $k \in \mathbb{N}^*$, the resulting sequence of iterates $(w_k)_{k \in \mathbb{N}}$ satisfies: $w_k = w_{k-1} - \gamma \mathcal{C}_k(g_k(w_{k-1}))$.*

And finally, the extension to the distributed case.

Algorithm 3 (Distributed compressed LMS) *In our motivating example, each client $i \in [N]$ observes at any step $k \in [K]$ an oracle $g_k^i(\cdot)$ on the gradient of the local objective function F_i, and applies a random compression mechanism $\mathcal{C}_k^i(\cdot)$. Thus, for any step-size $\gamma > 0$ and any $k \in \mathbb{N}^*$, the resulting sequence of iterates $(w_k)_{k \in \mathbb{N}}$ satisfies: $w_k = w_{k-1} - \frac{\gamma}{N} \sum_{i=1}^N \mathcal{C}_k^i(g_k^i(w_{k-1}))$ (we consider the randomization made on clients $(\mathcal{C}_k^i(\cdot))_{i \in \{1 \cdots, N\}}$ to be independent)*

Remark 3 *The analysis naturally covers any randomized postprocessing $\mathcal{C}_k^i(\cdot)$, beyond the compression case.*

Challenges, contributions and structure of the paper. Although there is abundant literature on the study of (LSA), the application to Algorithms 2 and 3 poses novel challenges. Especially, most analyses of LSA (Blum, 1954; Ljung, 1977; Ljung and Söderström, 1983) assume that the field ξ_k is linear (i.e. for any $z, z' \in \mathbb{R}^d$, $\xi_k(z) - \xi_k(z') = \xi_k(z - z')$, see Konda and Tsitsiklis, 2003; Benveniste et al., 2012; Leluc and Portier, 2022). More general non-asymptotic results on SA with a Lipschitz mean-field (i.e. SGD with a smooth objective) also assume that the noise-field is Lipschitz-in-squared-expectation i.e. for any $z, z' \in \mathbb{R}^d, \mathbb{E}[\|\xi_k(z) - \xi_k(z')\|^2] \leq C\|z - z'\|^2$ (Moulines and Bach, 2011; Bach, 2014; Dieuleveut et al., 2020; Gadat and Panloup, 2023). One major specificity and bottleneck in the case of compression is the fact that the resulting field **does not** satisfy such an assumption. The rest of the paper is thus organized as follows:

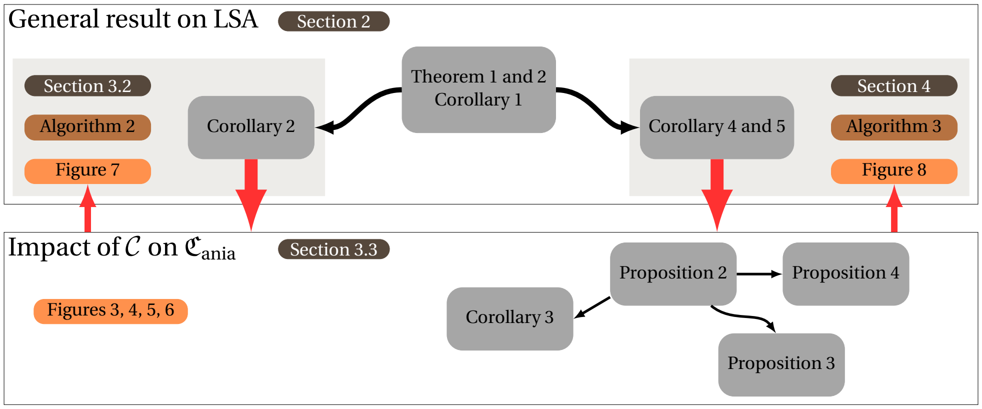


Figure 1: Flow chart summarizing our results.

1. In Section 2, we provide a non-asymptotic analysis of (LSA) under weak regularity assumptions of the noise field $(\xi_k)_k$. We show that the asymptotically dominant term depends on the covariance matrix $\mathfrak{C}_{\text{ania}}$ of the *additive noise induced by the algorithm*, as expected from the classical asymptotic literature (Polyak and Juditsky, 1992). The backbone results of our paper are Theorems 8 and 12 which generalize the results from Bach and Moulines (2013) for Algorithm 1. The limit convergence rate term scales with $\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}} H_F^{-1}\right)/K$, which highlights the interaction between the Hessian of the optimization problem H_F, and the additive noise's covariance $\mathfrak{C}_{\text{ania}}$.

2. In Section 3, we prove that assumptions made in Section 2 are valid for Algorithm 2 with classical compression schemes. Although this single-client case is a simple configuration, it enables to describe the impact of the compressor choice on the dependency between the features' covariance H (which is also the Hessian H_F of the optimization problem) and the additive noise's covariance $\mathfrak{C}_{\text{ania}}$. Contrary to Algorithm 1, for which the noise is said to be *structured*, i.e. the additive noise's covariance is proportional to the Hessian H_F, applying a random compression mechanism on the gradient breaks this structure. This phenomenon is noteworthy: for an ill-conditioned H_F, it may lead to a drastic increase in $\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}} H_F^{-1}\right)$ and thus, to a degradation in convergence. By calculating the additive noise's covariance for various compression mechanisms, we identify differences that classical literature was unable to capture.

3. In Section 4, we study the distributed Algorithm 3 with heterogeneous clients. We examine two different sources of heterogeneity for which we show that Theorems 8 and 12 remain valid. First, the case of heterogeneous features' covariances $(H_i)_{i=1}^N$ in Subsection 4.1; second, the case of heterogeneous local optimal points $(w_*^i)_{i=1}^N$ in Subsection 4.2.

These results are validated by numerical experiments which help to get an intuition of the underlying mechanisms. The code is provided on our GitHub repository: `https://github.com/philipco/structured_noise`. We summarize hereafter the structure of the paper in Figure 1.

Notations. We denote by $\preccurlyeq$ the order between self-adjoint operators, i.e., $A \preccurlyeq B$ if and only if $B - A$ is positive semi-definite (p.s.d.) and $A \mathrel{\widetilde{\preccurlyeq}} B$ if $A \preccurlyeq B$ and $A = B + O(\frac{1}{d})$. We denote by $A^{1/2}$ the p.s.d. square root of any symmetric p.s.d. matrix A. For two vectors

x, y in $\mathbb{R}^d$, the Kronecker product is defined as $x \otimes y := xy^\top$, the element-wise product is denoted as $x \odot y$, and the Euclidean norm is $\|x\|^2 := \sum_{i=1}^d x_i^2$. For any rectangular matrix A in $\mathbb{R}^{n\times m}$ s.t. $AA^\top$ is inversible, we denote $A^\dagger := A^\top(AA^\top)^{-1}$ the Moore–Penrose pseudo inverse. For x, y in $\mathbb{R}^d$, we use $x \wedge y$ for the minimum between two values, and $x \overset{\sim}{\leq} y$ if $x \leq y$ and $x = y + O(\frac{1}{d})$. For any sequence of vector $(x_k)_{k\in\{0,\dots,K\}}$ we denote $\overline{x}_{K-1} = \sum_{k=0}^{K-1} x_k/K$. We use e_i to denote the vector in $\mathbb{R}^d$ with zero everywhere except at coordinate i, and $\mathcal{O}_d(\mathbb{R})$ the group of orthogonal matrices. Finally, all random variables are defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, $\mathbb{E}$ is the expectation associated with the probability $\mathbb{P}$ and $\mathcal{A}$ is a σ-algebra. We define the set of probability distribution function $\mathcal{P}_M$ whose second moment is equal to M in $\mathbb{R}^{d\times d}$: $\mathcal{P}_M = \{\text{probability distribution } p_M \text{ over } \mathbb{R}^d \text{ s.t.}, \mathbb{E}_{\varepsilon\sim p_M}[\varepsilon^{\otimes 2}] = M\}$. Any such distribution p_M is indexed with its matrix of covariance.

2. Non asymptotic convergence result for (LSA)

2.1 Definition of the additive noise's covariance and assumptions on the random fields

For any k in $\mathbb{N}^*$, we define the additive noise ξ_k^{add} and the multiplicative noise $\xi_k^{\mathrm{mult}}(\cdot)$.

Definition 4 (Additive and multiplicative noise) *Under the setting of Definition 2, for any k in $\mathbb{N}^*$, we define:*

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

Remark 5 *Observe that $(\xi_k^{\mathrm{add}})_{k\in\mathbb{N}^*}$ is an i.i.d. sequence of random variables and $(\xi_k^{\mathrm{mult}})_{k\in\mathbb{N}^*}$ is an i.i.d. sequence of random field. The following assumptions, made for $k = 1$, are thus equivalently valid for any $k \geq 1$.*

Assumption 1 (Second moment) ξ_1^{add} *admits a second order moment. We note $\mathcal{A} \geq 0$ such that $\mathbb{E}[\|\xi_1^{\mathrm{add}}\|^2] \leq \mathcal{A}$.*

Assumption 1 and Remark 5 enable us to define the covariance of the additive noise induced by the algorithm.

Definition 6 (Additive noise's induced by the algorithm's covariance.) *Under the setting of Definition 2, we define the* additive noise's covariance *as the covariance of the additive noise: $\mathfrak{C}_{\mathrm{ania}} = \mathbb{E}[\xi_1^{\mathrm{add}} \otimes \xi_1^{\mathrm{add}}]$.*

Secondly, we state our assumptions on the multiplicative part of the noise, especially its regularity around 0 (note that $\xi_1^{\mathrm{mult}}(0) = 0$).

Assumption 2 (Second moment of the multiplicative noise) *There exist two constants $\mathcal{M}_1, \mathcal{M}_2 > 0$ such that, for any η in $\mathbb{R}^d$, the following hold:*

A2.1: $\mathbb{E}[\|\xi_1^{\mathrm{mult}}(\eta)\|^2] \leq 2\mathcal{M}_2 \|H_F^{1/2}\eta\|^2 + 4\mathcal{A}$.

A2.2: $\mathbb{E}[\|\xi_1^{\mathrm{mult}}(\eta)\|^2] \leq \mathcal{M}_1 \|H_F^{1/2}\eta\| + 3\mathcal{M}_2 \|H_F^{1/2}\eta\|^2$.

The main originality of this section is the analysis under Assumption 2.2. This Hölder-type condition will appear naturally for compression in Section 3. Up to our knowledge, (LSA) has not been analyzed under this particular condition.

Under these assumptions, asymptotic results from Polyak and Juditsky (1992) can be applied. Especially, we establish the asymptotic normality of $(\sqrt{K}\overline{\eta}_{K-1})_{K>0}$, with an asymptotic variance equal to $H_F^{-1}\mathfrak{C}_{\mathrm{ania}}H_F^{-1}$.

Proposition 7 (CLT for (LSA)**)** *Under Assumptions 1 and 2, consider a sequence* $(w_k)_{k\in\mathbb{N}^*}$ *produced in the setting of Definition 2 for a step-size* $(\gamma_k)_{k\in\mathbb{N}^*}$ *s.t.* $\gamma_k = k^{-\alpha}$, $\alpha \in]0,1[$. *Then* $(\sqrt{K}\overline{\eta}_{K-1})_{K>0}$ *is asymptotically normal and converge in distribution to* $\mathcal{N}(0, H_F^{-1}\mathfrak{C}_{\mathrm{ania}}H_F^{-1})$.

The proof of this result is almost straightforward and is recalled in Appendix A.4. In the following, we establish non-asymptotic results in Theorems 8 and 12, that highlight the impact of Assumption 2.2.

2.2 Convergence rates for (LSA), general case

In this section, we present non-asymptotic convergence rates for (LSA) under the assumptions above. These results build upon the work of Bach and Moulines (2013). Our first result is the main result, under the Hölder assumption on the noise field, it is demonstrated in Appendix B.

Theorem 8 (Non-linear multiplicative noise) *Under Assumptions 1 and 2, consider a sequence* $(w_k)_{k\in\mathbb{N}^*}$ *produced in the setting of Definition 2 for a constant step-size* γ *such that* $\gamma(R_F^2 + 2\mathcal{M}_2) \leq 1/2$. *Then for any horizon* K, *we have:*

$$\mathbb{E}[F(\overline{w}_{K-1}) - F(w_*)] \leq \frac{1}{2K}\left(\frac{\|H_F^{-1/2}\eta_0\|}{\gamma\sqrt{K}} \wedge \frac{\|\eta_0\|}{\sqrt{\gamma}} + \sqrt{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)} + (10\mathcal{A}\gamma)^{1/4}\sqrt{\mathcal{M}_1\mu^{-1}} + (30\mathcal{A}\gamma)^{1/2}\sqrt{\mathcal{M}_2\mu^{-1}}\right)^2 .$$

The first two terms of the RHS correspond respectively to the impact of the initial condition η_0 and the impact of the additive noise. The dependency on these two terms is similar to the one established in Bach and Moulines (2013) in the case of LMS. Note that following Defossez and Bach (2015), we improve the dependency on the initial condition to $\frac{\|\eta_0\|^2}{\gamma K} \wedge \frac{\|H_F^{-1/2}\eta_0\|^2}{\gamma^2K^2}$. Regarding the noise term, the dependency on $\frac{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)}{2K}$ corresponds to the classical asymptotic noise term in CLT for Stochastic Approximation (e.g., Delyon, 1996; Duflo, 1997; Györfi and Walk, 1996). In fact, for a sequence of step sizes γ_t decreasing to zero, we recover the variance from Proposition 7. Remark that in (Bach and Moulines, 2013) and several follow up works, the algorithm under consideration is LMS (Algorithm 1, which enables to ensure that $\mathfrak{C}_{\mathrm{ania}} \preccurlyeq \sigma^2 H_F$: the variance term thus scales as $\sigma^2 d/K$. On the contrary, Algorithms 2 and 3 do not always satisfy $\mathfrak{C}_{\mathrm{ania}} \preccurlyeq \sigma^2 H_F$: in such case, $\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)$ may scale as $1/\mu$.

The third and fourth term, that scale respectively as $\sqrt{\gamma}/K$ and γ/K, are asymptotically negligible for $\gamma = o(1)$. Those term are proportional to the Hölder-regularity constants

$\mathcal{M}_1, \mathcal{M}_2$, and also increase with μ^{-1}. The dominant term is $\frac{\mathcal{M}_1\sqrt{10\mathcal{A}\gamma}}{\mu K}$. Interestingly, when γ is constant (not decreasing with K), then the limit variance of the algorithm is affected. Moreover, contrary to (Bach and Moulines, 2013), we do not recover a convergence rate independent of μ. This dependency is un-avoidable as the multiplicative noise is only controlled around w_*: without strong-convexity, the iterates may not converge to w_*. While these additional terms in the variance may be considered as a drawback, it can be mitigated by taking a step-size γ proportional to $1/K^\alpha$ with $\alpha > 0$ small (γ is horizon dependent, but constant).

Corollary 9 *Under the assumptions of Theorem 8, with $\gamma = 1/K^\alpha$, and $\alpha \in]0, 1/2[$, we have:*

$$\mathbb{E}[F(\overline{w}_{K-1}) - F(w_*)] \leq \frac{60}{K}\left(\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}} H_F^{-1}\right) + \frac{\|H_F^{-1/2}\eta_0\|^2}{K^{(1-2\alpha)}} + \frac{\mathcal{M}_1\sqrt{\mathcal{A}}}{\mu K^{\alpha/2}} + \frac{\mathcal{M}_2\mathcal{A}}{\mu K^\alpha}\right).$$

The decrease of the second order terms is then optimized for $\alpha = 2/5$. To highlight the impact of the non-linearity in compression schemes, we provide for comparison the result for a linear multiplicative noise.

2.3 Convergence rates for (LSA), linear case

Alternatively, to cover the particular case of a linear multiplicative noise (e.g., to recover LMS or projection-based compressed LMS) we make the following stronger hypothesis:

Assumption 3 *The multiplicative noise is linear i.e. there exists a random matrix Ξ_1 in $\mathbb{R}^{d\times d}$ s.t. for any η in $\mathbb{R}^d$, we have a.s. $\xi_1^{\text{mult}}(\eta) = \Xi_1\eta$. Moreover $\mathbb{E}[\|\xi_1^{\text{mult}}(\eta)\|^2] \leq \mathcal{M}_2\|H_F^{1/2}\eta\|^2$.*

Remark 10 *Note that Ξ_1 is not necessarily symmetric (in Algorithms 2 and 3, this results from the compression).*

In addition to Assumption 3, in the case of linear multiplicative noise, we also consider the following assumption.

Assumption 4 *The following hold.*

A4.1: *There exists a constant*[3] $\text{Ш}_{\text{add}} > 0$ *s.t.* $\mathfrak{C}_{\text{ania}} \preccurlyeq \text{Ш}_{\text{add}} H_F$.

A4.2: *There exists a constant* $\text{Ш}_{\text{mult}} > 0$, *such that* $\mathbb{E}\left[\Xi_1\Xi_1^\top\right] \preccurlyeq \text{Ш}_{\text{mult}} H_F$.

Remark 11 (Link between Assumptions 1, 2 and 4) *Assumption 1 (resp. Assumption 2) corresponds to an assumption on the second order moment of the additive noise (resp. multiplicative), while Assumption 4.1 (resp. Assumption 4.2) is a (stronger) assumption on its covariance.*

3. This letter Ш is the Russian upper letter "sha".

Theorem 12 (Linear multiplicative noise) *Under Assumptions 1, 3 and 4, i.e., with a linear multiplicative noise. Consider a sequence* $(w_k)_{k\in\mathbb{N}^*}$ *produced in the setting of Definition 2, for a constant step-size* γ *such that* $\gamma(R_F^2+\mathcal{M}_2)\le 1$ *and* $4Ш_{\mathrm{mult}}\gamma\le 1$*. Then for any horizon* K*, we have*

$$\mathbb{E}[F(\overline{w}_{K-1})-F(w_*)]\le\frac{1}{2K}\left(\frac{\|\eta_0\|}{\sqrt{\gamma}}+\sqrt{\operatorname{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)+2\sqrt{\gamma dШ_{\mathrm{add}}Ш_{\mathrm{mult}}}}\right)^2 .$$

Theorem 12 generalizes Theorem 1 from Bach and Moulines (2013). It also highlights the impact of additive noise's covariance, and the comparison between Theorem 8 and Theorem 12 shows the advantage of linear compression schemes. Indeed the variance scales as $K^{-1}(\operatorname{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)+4\gamma dШ_{\mathrm{add}}Ш_{\mathrm{mult}})$. As before, the first term $\operatorname{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)$ corresponds to the asymptotic variance given in Proposition 7, and the second term is negligible: (i) for all $4Ш_{\mathrm{mult}}\gamma\le 1$ it can be upper bounded by $dШ_{\mathrm{add}}$, and for LMS (see Bach and Moulines, 2013), the variance term is $\operatorname{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)=d\sigma^2$, which is thus at least as large, (ii) it scales with γ thus is asymptotically negligible as γ tends to 0. Overall, depending on $\mathfrak{C}_{\mathrm{ania}}$, the algorithm may or may not suffer from the lack of strong-convexity (μ tending to 0). More precisely, in the case of linear multiplicative noise, we can obtain a $O(K^{-1})$ rate independent of μ if and only if $\mathfrak{C}_{\mathrm{ania}}\preccurlyeq aH_F$, with a in $\mathbb{R}$. The proof of Theorem 12 is given in Appendix C, and follows the line of proof of Bach and Moulines (2013).

Conclusion: we established rates for (LSA) for both the Hölder-noise case and the linear noise case. In the former, convergence requires strong convexity while in the latter, we can achieve $O(K^{-1})$ for $\mathfrak{C}_{\mathrm{ania}}\preccurlyeq aH_F$. In both cases, the dominant term for an optimal choice of γ scales as $\frac{\operatorname{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)}{K}$.

In the following section, we turn to the analysis of Algorithm 2: we show how the choice of the compression impacts both the linearity of the noise and the structure of $\mathfrak{C}_{\mathrm{ania}}$.

3. Application to Algorithm 2: compressed LSR on a single worker

In this section, we analyze Algorithm 2, i.e. compressed LSR. In Subsection 3.1, we introduce the compression operators of interest and verify in Subsection 3.2 that Theorems 8 and 12 can be applied. Then, in Subsection 3.3, we provide explicit formulas of $\operatorname{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H^{-1}\right)$ for various compression schemes. Finally, in Subsection 3.4, we validate our findings with numerical experiments.

3.1 Compression operators

Our analysis applies to most unbiased compression operators.

Definition 13 (Compression operators) *Let* $z\in\mathbb{R}^d$*.*

1. 1***-quantization*** *is defined as* $\mathcal{C}_{\mathrm{q}}(z):=\|z\|\operatorname{sign}(z)\odot\chi$ *with* $\chi\sim\otimes_{i=1}^d(\mathrm{Bern}(|z_i|/\|z\|_2))$.
2. ***Stabilized*** 1***-quantization*** *is defined as* $\mathcal{C}_{\mathrm{sq}}(z):=U^\top\mathcal{C}_q(Uz)$*, with* $U\in\mathrm{Unif}(\mathcal{O}_d)$*.*
3. ***Rand-****h* *is defined as* $\mathcal{C}_{\mathrm{rd}h}(z):=\frac{d}{h}B(S)\odot z$ *with* $S\sim\mathrm{Unif}(\mathcal{P}_h([d]))$ *and* $B(S)_i=\mathbb{1}_{i\in S}$*.*
4. ***Sparsification*** *is defined as* $\mathcal{C}_{\mathrm{s}}(z):=\frac{1}{p}B\odot z\in\mathbb{R}^d$ *with* $B\sim\otimes_{i=1}^d\left(\mathrm{Bern}(p)\right)$ *.*

5. ***Partial participation*** *is defined* $\mathcal{C}_{\mathrm{PP}}(z) := \frac{b_0}{p} z$ *with* $b_0 \sim \mathrm{Bern}(p)$*.*
6. ***Random Projection****, also referred to as* sketching*, is defined as* $\mathcal{C}_\Phi(z) := \frac{1}{p}\Phi^\dagger \Phi z$*, where* $h \ll d \in \mathbb{N}$*,* $p = h/d$ *and* $\Phi \in \mathbb{R}^{h\times d}$ *is a random projection matrix onto a lower-dimension space (Vempala, 2005; Li et al., 2006). In the following, we consider Gaussian projection, where each element* $i, j \in [\![1, h]\!] \times [\![1, d]\!]$ *follows an independent zero-centered normal distribution.*

We refer to the introduction for related work on compression. Operators $\mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{sq}}$ are quantization-based schemes while $\mathcal{C}_{\mathrm{rd1}}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\mathrm{PP}}, \mathcal{C}_\Phi$ are projection-based. Indeed sparsification can be seen as a random projection (for $h \ll d$, $p = h/d$ and h randomly sampled coordinates $\mathcal{I}$ from $[\![1, d]\!]$ such that for any $i \in \mathcal{I}$, the i^{th} lines of Φ are equal to $e_i \in \mathbb{R}^d$, and equal to zero otherwise). For $\mathcal{C}_{\mathrm{PP}}$, the motivation is distributed settings, in which the intermittent availability of clients prevents them from systematically participating in the training. This can be modeled through *partial participation*: clients only participate in a fraction p of the training steps. In theoretical analyses, this can be handled as a compression scheme $\mathcal{C}_{\mathrm{PP}}$, in which the compression of a vector z is either z/p or 0. Observe that in the centralized case, this is slightly artificial as it actually means that no update is performed at most steps and that the step-size is scaled at the other steps. Finally, we denote $\mathcal{C}_{\mathrm{I}_d} : z \in \mathbb{R}^d \mapsto z$ the operator that does not carry out any compression.

Remark 14 *The analysis of random projection is related to Random features (Rahimi and Recht, 2008), usually used for Kernel learning in infinite dimensions. Nyström method (introduced by Kumar et al., 2009) is another similar technique of compression often used in this setting, it consists of removing a subset* $\mathcal{S} \subset \{1, \cdots, d\}$ *of lines and columns in the kernel matrix* K*. Both techniques have been extensively studied in the context of linear and non-linear kernel learning (Rudi et al., 2015, 2017; Rudi and Rosasco, 2017; Lin and Rosasco, 2017). Recently, the combination of SGD and random features has been analyzed by Carratino et al. (2018). However, their results cannot be directly applied to our setting for two reasons. Firstly, their analysis is for infinite dimensions, where they obtain a* $O(1/\sqrt{K})$ *rate of convergence. Secondly, the compressions used in their approach are not independent at each iteration.*

Remark 15 *Diffusion LMS (i.e. distributed learning without a central server) has also been studied from the perspective of low-cost training by Arablouei et al. (2015); Harrane et al. (2018), but using only clients' partial participation or sparsification. Contrary to our work they use biased compression and an adaptive correction step to compensate for the induced error. They provide results guarantying asymptotic convergence (Harrane et al., 2018, see Equations (28)-(37)).*

3.2 Applicability of the results on (LSA) from Section 2

We first show that our results from Section 2 can be applied for Algorithm 2 with a random compression operator $\mathcal{C}$, in the case of Model 2.

Lemma 16 *For any compressor* $\mathcal{C} \in \{\mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{sq}}, \mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_\Phi, \mathcal{C}_{\mathrm{PP}}\}$*, there exists constants* $\omega, \Omega \in \mathbb{R}_+^*$*, such that the random operator* $\mathcal{C}$ *satisfies the following properties for all* $z, z' \in \mathbb{R}^d$*.*

L.1: $\mathbb{E}[\mathcal{C}(z)] = z$ *and* $\mathbb{E}[\|\mathcal{C}(z)-z\|^2] \leq \omega \|z\|^2$ *(unbiasedness and variance relatively bounded),*
L.2: $\mathbb{E}[\|\mathcal{C}(z) - \mathcal{C}(z')\|^2] \leq \Omega \min(\|z\|, \|z'\|)\|z - z'\| + 3(\omega + 1)\|z - z'\|^2$ *(Hölder-type bound),*

with $\omega = \sqrt{d}$ *and* $\Omega = 12\sqrt{d}$ *(resp.* $\omega = (1-p)/p$ *and* $\Omega = 0$*) for* $\mathcal{C}_{\mathrm{q}}$ *and* $\mathcal{C}_{\mathrm{sq}}$ *(resp.* $\mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_{\mathrm{s}}$, $\mathcal{C}_{\Phi}, \mathcal{C}_{\mathrm{PP}}$*).*

We note $\mathbb{C}$ the set of unbiased compressors verifying Lemma 16. Item L.1 is frequently established in the literature and corresponds to the worst-case assumption, see the introduction for references. On the other hand, Item L.2 is the Hölder-type bound, which is not used in the literature up to our knowledge. The expected squared distance between the compression of two nearby points scales with the *non-squared* norm of the distance. Moreover, the distance is multiplied by an unavoidable coefficient scaling with z, z'. Remark that in Item L.2, we assume the compression randomness to be the same for the compression of z and z': formally, we control $\mathcal{W}_2(\mathcal{C}(z), \mathcal{C}(z'))^2$, with $\mathcal{W}_2$ the Wassertein-2 distance. This lemma is demonstrated in Appendix E.1.

Remark 17 *For a given* ω*, note that the* communication cost c *for quantization-based and projection-based compressors is not always equivalent. For* 1*-quantization we have* $c \approx \frac{3}{2}\sqrt{d}\log_2 d + 32$ *while for projection-based we have* $c \approx 32\sqrt{d}$*, for* $\sqrt{d}$*-quantization we have* $c \approx 3d + 32$ *while for projection-based we have* $c = 16d$*.*

Lemma 16 enables to show that Theorems 8 and 12, and Algorithm 2 are valid in the context of Model 2.

Corollary 18 *Consider Algorithm 2 in the context of Model 2, with a compressor* $\mathcal{C} \in \{\mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{sq}}, \mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\Phi}, \mathcal{C}_{\mathrm{PP}}\}$*. With Lemma 16 above, Assumptions 1 and 2 on the resulting random field* $(\xi_k)_{k\in\mathbb{N}^*}$ *are valid, with in particular* $H_F = H$*,* $R_F^2 = R^2$*,* $\mathcal{A} = (\omega + 1)R^2\sigma^2$*,* $\mathcal{M}_2 = (\omega + 1)R^2$*,* $\mathcal{M}_1 = \Omega R^2 \sigma$*. Therefore, it follows that Theorem 8 holds.*

Moreover for any linear compressor $\mathcal{C} \in \{\mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\Phi}, \mathcal{C}_{\mathrm{PP}}\}$*, under Remark 1, we also have that Assumptions 3 and 4 are valid with* $\text{Ш}_{\mathrm{add}} = \sigma^2 \text{Ш}_H$ *and* $\text{Ш}_{\mathrm{mult}} = R^2 \text{Ш}_H$*, with* Ш_H *given below. Therefore, it follows that Theorem 12 holds.*

Compressor	$\mathcal{C}_{\mathrm{rd}h}$	$\mathcal{C}_{\mathrm{s}}$	$\mathcal{C}_{\mathrm{PP}}$	$\mathcal{C}_{\Phi}$
Ш_H	$\frac{h-1}{p(d-1)} + (1 - \frac{h-1}{d-1})\frac{\tau}{p}$	$1 + \frac{(1-p)\tau}{p}$	$\frac{1}{p}$	$\frac{\alpha-\beta}{p} + \frac{\beta\tau}{p}$
Ш_H *(if H diagonal)*	$\frac{1}{p}$	$\frac{1}{p}$	$\frac{1}{p}$	$\frac{\alpha-\beta}{p} + \frac{\beta\tau}{p}$

Where $p = h/d$*,* $\tau = \mathrm{Tr}(H)/\mu$*, and for sketching* $\alpha = \frac{h+2}{d+2}$ *and* $\beta = \frac{d-h}{(d-1)(d+2)}$*.*

This corollary is proved in Appendix D. We observe that a first difference in terms of convergence exists between quantization-based compression and projection-based: for the former, *only* Theorem 8 can be applied and the lower-order terms always have a *poorer dependency on* μ while for the latter, Theorem 12 is applicable and lower-order terms do not necessarily depends on μ. Indeed, the constants Ш_H do not depend on μ for $\mathcal{C}_{\mathrm{PP}}$, and for $\mathcal{C}_{\mathrm{rd1}}, \mathcal{C}_{\mathrm{s}}$, *when* the features' covariance H is diagonal. On the contrary, there is always a dependency on μ for $\mathcal{C}_{\Phi}$, and for $\mathcal{C}_{\mathrm{rd1}}, \mathcal{C}_{\mathrm{s}}$ when H is not diagonal. In practice, this means that, among projection-based compressors, regarding lower-order terms, the convergence is expected to be slower for random Gaussian projection.

We now turn to the analysis of the impact of the choice of the compression on the dominant asymptotic term $\mathrm{Tr}\left(H_F^{-1}\mathfrak{C}_{\mathrm{ania}}\right)$.

3.3 Impact of the compression on the additive noise covariance

In this section, we illustrate how distinct compressors lead to different covariances for the additive noise. This shows how $\mathrm{Tr}\left(H_F^{-1}\mathfrak{C}_{\mathrm{ania}}\right)$ is impacted by the choice of a compressor.

First recall that for Algorithm 2 in the context of Model 2, with any compressor $\mathcal{C}$, the additive noise writes for any $k \in [K]$, as:

$$\xi_k^{\mathrm{add}} \overset{\text{def. } 4}{=} \xi_k(0) \overset{\text{algo } 2}{=} \nabla F(w_*) - \mathcal{C}_k(g_k(w_*)) \overset{\text{eq. } 2}{=} -\mathcal{C}_k((\langle x_k, w_*\rangle - y_k)x_k) \overset{\text{model } 2}{=} \mathcal{C}_k(\varepsilon_k x_k)\,.$$

Also recall that $\mathfrak{C}_{\mathrm{ania}}$ is defined as $\mathfrak{C}_{\mathrm{ania}} := \mathbb{E}[(\xi_k^{\mathrm{add}})^{\otimes 2}] = \mathbb{E}[\mathcal{C}(\varepsilon_k x_k)^{\otimes 2}]$. Moreover, note that $\mathcal{C}(\varepsilon_k x_k) \overset{\text{a.s.}}{=} \varepsilon_k \mathcal{C}(x_k)$ for all operators under consideration (this is immediate for linear operators and results from the scaling for quantization-based ones). Consequently

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

as $\mathbb{E}[\varepsilon_k^2 | x_k] = \sigma^2$. Ultimately, we have to study the covariance of $\mathcal{C}(x_k)$, for x_k a random variable with second-moment H.

We thus generically study the covariance of $\mathcal{C}(E)$, for E a random vector with distribution p_M with second moment[4] $\mathbb{E}[E^{\otimes 2}] = M$.

Definition 19 (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* $p_M \in \mathcal{P}_M$*,*

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

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

Using a compressor $\mathcal{C} \in \mathbb{C}$, we therefore have by Equation (3):

$$\mathfrak{C}_{\mathrm{ania}} = \sigma^2 \mathfrak{C}(\mathcal{C}, p_H), \tag{4}$$

where p_H is the marginal distribution of x_k (for any k).

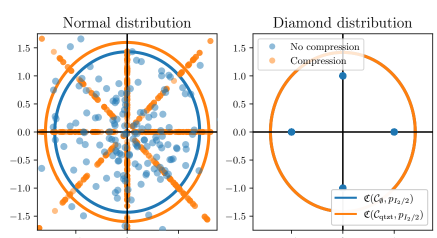


Figure 2: Illustration of Remark 20

4. Remark that we do not assume $\mathbb{E}[E] = 0$. Indeed, all computations only depend on the *second-order moment* M of E, not on its variance (and the convergence depends of the second-*order moment* H of x, not its variance). It is clear, that $\mathbb{E}[\mathcal{C}(E)^{\otimes 2}]$ does not depend on the fact that E is centered: indeed, for R a Rademacher $1/2$ independent of E, we have $\mathbb{E}[\mathcal{C}(E)^{\otimes 2}] = \mathbb{E}[R^2]\mathbb{E}[\mathcal{C}(E)^{\otimes 2}] \overset{\perp}{=} \mathbb{E}[(R\mathcal{C}(E))^{\otimes 2}] = \mathbb{E}[\mathcal{C}(RE)^{\otimes 2}]$ and RE is (1) centered (2) has the same second-moment as E. Remark that centering the covariates before learning does impact H: indeed $H = \mathbb{E}[(x)^{\otimes 2}] = \mathbb{E}[(x - \mathbb{E}[X])^{\otimes 2}] + (\mathbb{E}[X])^{\otimes 2})$. Centering subtracts $(\mathbb{E}[X])^{\otimes 2}$ to the second moment, which is a rank-1 matrix, typically does not affect the smallest eigenvalue, but it can affect the top-eigenvalue.

Remark 20 (Dependence on p_M, not only M) *Note that, for $\mathcal{C} = \mathcal{C}_{\mathrm{q}}$, there exist two distributions p_M, p'_M with the same covariance M, such that $\mathfrak{C}(\mathcal{C}, p_M) \neq \mathfrak{C}(\mathcal{C}, p'_M)$. This is why we cannot simply denote $\mathfrak{C}(\mathcal{C}, M)$.*

Indeed, consider $d = 2$ and (1) a normal distribution $E_1 \sim \mathcal{N}(0, I_2/2)$, vs (2) a *diamond* distribution $E_2 \sim \mathbb{P}_\diamond$, such that $\mathbb{P}_\diamond\{(1,0)\} = \mathbb{P}_\diamond\{(-1,0)\} = \mathbb{P}_\diamond\{(0,1)\} = \mathbb{P}_\diamond\{(0,-1)\} = 1/4$, and thus $\mathrm{Cov}\,[E_1] = \mathrm{Cov}\,[E_2] = I_2/2$. Then $\mathrm{Cov}\,[E_1] \prec \mathrm{Cov}\,[\mathcal{C}_{\mathrm{q}}(E_1)]$, but $\mathcal{C}_{\mathrm{q}}(E_2) \overset{\text{a.s.}}{=} E_2$ thus $\mathrm{Cov}\,[E_2] = \mathrm{Cov}\,[\mathcal{C}_{\mathrm{q}}(E_2)]$. We illustrate this on Figure 2: we represent E_i in blue and $\mathcal{C}_{\mathrm{q}}(E_i)$ in orange for $i = 1$ (left) and $i = 2$ (right). We also represent the covariance matrices by plotting the ellipses $\mathcal{E}_{\mathrm{Cov}[E_i]}$ and $\mathcal{E}_{\mathrm{Cov}[\mathcal{C}_{\mathrm{q}}(E_i)]}$, where $\mathcal{E}_M = \{x \in \mathbb{R}^d, x^\top M^{-1} x = 4\}$ (see Definition S34)[5].

We now compute for the compression operators, the value or an upper bound on $\mathfrak{C}(\mathcal{C}, p_H)$.

Proposition 21 (Compression and covariance) *The following formulas hold:*

$$
\begin{aligned}
\mathfrak{C}(\mathcal{C}_{\mathrm{I}_d}, p_M) &= M \\
\mathfrak{C}(\mathcal{C}_q, p_M) &\preccurlyeq \widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) := M + \sqrt{\mathrm{Tr}\,(M)}\sqrt{\mathrm{Diag}\,(M)} - \mathrm{Diag}\,(M) \\
&\quad \textit{(with equality if } \|E\| \textit{ is a.s. constant under } p_M\textit{)} \\
\mathfrak{C}(\mathcal{C}_s, p_M) &= M + (1-p)p^{-1}\mathrm{Diag}\,(M) \\
\mathfrak{C}(\mathcal{C}_\Phi, p_M) &= p^{-1}\left((\tfrac{h+1}{d+2} + \delta_{hd})M + \left(1 - \tfrac{h-1}{d-1}\right) \tfrac{\mathrm{Tr}(M)}{d+2} \mathrm{I}_d \right), \ \textit{with } \delta_{hd} = \tfrac{h-1}{(d-1)(d+2)} = O(\tfrac{1}{d}) \\
\mathfrak{C}(\mathcal{C}_{\mathrm{rd}h}, p_M) &= p^{-1}\left(\tfrac{h-1}{d-1} M + \left(1 - \tfrac{h-1}{d-1}\right) \mathrm{Diag}\,(M) \right) \\
\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) &= p^{-1} M\,.
\end{aligned}
$$

Conclusion and interpretation. Most compression operators induce *both* a *structured* noise (Flammarion and Bach, 2015) which covariance scales with H and an *unstructured* noise, which covariance scales with $\mathrm{Diag}\,(H)$ or I_d—thus corresponding to an *isotropic* noise.

From the convergence standpoint, the asymptotic convergence rate scales with the trace $\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}} H^{-1}\right) = \sigma^2 \mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}, p_H) H^{-1}\right)$. Therefore, the un-structured part in the noise is problematic as $\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}} H^{-1})\right)$ will strongly depends on the smallest eigenvalue μ. This comes from the fact that the compression induces a significant noise in directions in which the Hessian curvature is very limited (thus directions onto which the contraction towards the optimum in the algorithm is weak).

A particular case is when H is diagonal (e.g. the features are *centered* and *independent*), we get the following corollary.

Corollary 22 (Compression and covariance, diagonal case) *If M is diagonal, then Proposition 21 is simplified to the following (with the same δ_{hd}):*

$$
\begin{aligned}
\mathfrak{C}(\mathcal{C}_{\mathrm{I}_d}, p_M) &= M & \mathfrak{C}(\mathcal{C}_\Phi, p_M) &= p^{-1}\left((\tfrac{h+1}{d+2} + \delta_{hd})M + (1 - \tfrac{h-1}{d-1}) \tfrac{\mathrm{Tr}(M)}{d+2} \mathrm{I}_d \right) \\
\mathfrak{C}(\mathcal{C}_q, p_M) &\preccurlyeq \sqrt{\mathrm{Tr}\,(M)}\sqrt{M} & \mathfrak{C}(\mathcal{C}_{\mathrm{rd}h}, p_M) &= p^{-1} M \\
\mathfrak{C}(\mathcal{C}_s, p_M) &= p^{-1} M & \mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) &= p^{-1} M.
\end{aligned}
$$

5. The constant 4 is chosen so that for Gaussian distributions, the expected fraction of points within the ellipse is $86,4\% \simeq 1 - F_{\chi^2(2)}(4)$

Remark 23 (Composition of compressors) *For all compression schemes but $\mathcal{C}_{\mathrm{q}}$, we observe that $\mathfrak{C}(\mathcal{C}, p_M)$ is a function of M, which complements Remark 20. In that particular case, we can then denote $\mathfrak{C}(C, M)$. This means that the lemma can be extended to any composition of compression schemes, for example to compute $\mathfrak{C}(C_1 \circ C_2, M) = \mathfrak{C}(C_1, \mathfrak{C}(C_2, M))$.*

From Proposition 21 and Corollary 22 we can deduce certain generic comparisons between the asymptotic convergence rates, depending on the compression operator (for compression operators having the same variance bound). They are proven in Appendix E.3. In the following, for any $a, b \in \mathbb{R}$, we use the notation $a \overset{\sim}{\leq} b$, to denote a *systematic inequality* (i.e., $a \leq b$) with a negligible difference as $d \to \infty$ (i.e., $a = b + O(1/d)$), and similarly for any two symmetric matrices $A, B \in \mathcal{S}_d(\mathbb{R})$, $A \overset{\sim}{\preccurlyeq} B$, for $A \preccurlyeq B$ and $A = B + O(1/d)$ as $d \to \infty$.

Proposition 24 (Comparison between $\mathcal{C}_{\mathrm{PP}}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_{\Phi}$, $\omega = d/h - 1$) *We consider $\mathcal{C} \in \{\mathcal{C}_{\mathrm{PP}}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_{\Phi}\}$ with $p = h/d$, such that $\mathcal{C}$ always satisfies Lemma 16 with $\omega = d/h - 1$. For any matrix $M \in \mathbb{R}^{d \times d}$:*

1. *If M is diagonal, then:*

   - $\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) = \mathfrak{C}(\mathcal{C}_{\mathrm{s}}, p_M) = \mathfrak{C}(\mathcal{C}_{\mathrm{rd}h}, p_M) = \frac{d}{h} M$,
   - $\mathrm{Tr}\big(\mathfrak{C}(\mathcal{C}_{\mathrm{PP/s/rd}h}, p_M) M^{-1}\big) \leq \mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\Phi}, p_M) M^{-1}\right)$.

   This means that the asymptotic convergence rate does not depend on the choice of the compressor between $\mathcal{C}_{\mathrm{PP}}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\mathrm{rd1}}$ in the diagonal case.

2. *Moreover, for any matrix M with a* constant diagonal *(e.g., we standardize*[6] *the data in the pre-processing step, such that* $\mathrm{Diag}(M) = \mathrm{I}_d$*), we have:*

$$\mathrm{Tr}(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) M^{-1}) \leq \mathrm{Tr}(\mathfrak{C}(\mathcal{C}_{\Phi}, p_M) M^{-1}) \leq \mathrm{Tr}(\mathfrak{C}(\mathcal{C}_s, p_M) M^{-1}) \leq \mathrm{Tr}(\mathfrak{C}(\mathcal{C}_{\mathrm{rd}h}, p_M) M^{-1}),$$

   with strict inequalities if M is not proportional to I_d. This means that we expect the asymptotic convergence rate to be faster for PP than Sparsification, Sketching, or Rand-h (illustrated in experiments).

In the next proposition, we compare compressors $\mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\mathrm{PP}}$ to $\mathcal{C}_{\mathrm{q}}$ for equal $\omega = \sqrt{d}$ (we exclude $\mathcal{C}_{\mathrm{rd}h}$ and $\mathcal{C}_{\Phi}$ for which h must be an integer).

Proposition 25 (Comparison between $\mathcal{C}_{\mathrm{PP}}, \mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{s}}$, $\omega = \sqrt{d}$) *We consider that $\mathcal{C}$ is in $\{\mathcal{C}_{\mathrm{PP}}, \mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{s}}\}$ with $p = (\sqrt{d} + 1)^{-1}$, such that $\mathcal{C}$ always satisfies Lemma 16 with $\omega = \sqrt{d}$.*

1. *For any symmetric matrix M diagonal, we have:*

$$\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) M^{-1}\right) = \mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_s, p_M) M^{-1}\right) \overset{\textit{possib.} \ll}{\leq} \left(1 + \frac{1}{\sqrt{d}}\right) \mathrm{Tr}\left(\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) M^{-1}\right).$$

2. *If M is not necessarily diagonal but with a* constant diagonal *(e.g., after standardization), then*

6. That means we center and rescale to get a variance of one for each feature.

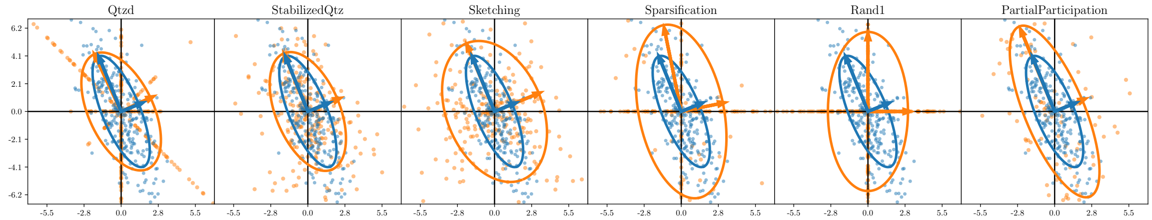


Figure 3: H not diagonal. Scatter plot of $(x_k)_{i=1}^K$/ $(\mathcal{C}(x_k))_{i=1}^K$ with its ellipse $\mathcal{E}_{\mathrm{Cov}[x_k]}/\mathcal{E}_{\mathrm{Cov}[\mathcal{C}(x_k)]}$.

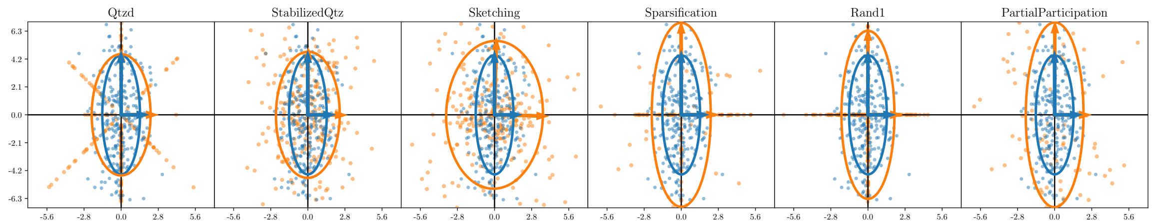


Figure 4: H diagonal. Scatter plot of $(x_k)_{i=1}^K/(\mathcal{C}(x_k))_{i=1}^K$ with its ellipse $\mathcal{E}_{\mathrm{Cov}[x_k]}/\mathcal{E}_{\mathrm{Cov}[\mathcal{C}(x_k)]}$.

- $\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) \preccurlyeq \mathfrak{C}(\mathcal{C}_{\mathrm{s}}, p_M)$
- $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) M^{-1}\right) \leq \left(1 + \frac{1}{\sqrt{d}}\right) \mathrm{Tr}\left(\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) M^{-1}\right)$

This means that sparsification is expected to always result in a poorer asymptotic convergence rate than quantization. Moreover, the upper bound on the covariance $\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M)$ *for quantization itself leads to a worst bound than for PP.*[7]

We now propose a detailed illustration of the results of Proposition 21 and Corollary 22, first in a low-dimensional setting ($d = 2$) and then in higher dimension on synthetic and real datasets.

3.3.1 Illustration of Proposition 21 and Corollary 22 in dimension 2.

In order to build intuition, we illustrate Proposition 21 and Corollary 22 in Figures 3 and 4, showing how compression affects the additive noise covariance, in a simple 2-dimensional case, for both a non-diagonal matrix M (Figure 3) and a diagonal one (Figure 4).

More specifically, we consider features $(x_k)_{k\in[K]}$ sampled from $\mathcal{N}(0, M)$ where $M = QDQ$, $D = \mathrm{Diag}\,(1, 10)$ and Q is rotation matrix with angle $\pi/8$ (resp. 0) in Figure 3 (resp. 4). We represent the values of x_k and $\mathcal{C}(x_k)$, unit-ellipses of the corresponding covariance matrices $\mathcal{E}_{\mathrm{Cov}[x_k]}$ and $\mathcal{E}_{\mathrm{Cov}[\mathcal{C}(x_k)]}$ (see Definition S34—recall that $\mathcal{E}_{\mathrm{Cov}[x_k]} \subset \mathcal{E}_{\mathrm{Cov}[\mathcal{C}(x_k)]} \Leftrightarrow \mathrm{Cov}\,[x_k] \preccurlyeq \mathrm{Cov}\,[\mathcal{C}(x_k)]$), as well as their two eigenvectors; we take $p = (1 + \sqrt{d})^{-1} = 0.41$, hence for $\mathcal{C} \in \{\mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{sq}}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\mathrm{PP}}\}$ we have $\omega = 1.41$ but for sketching and rand-1, we have $p = 1/2$ and $\omega = (1 - p)/p = 1$.

We make the following observations:

[Qtz] For quantization and stabilized quantization, in the non-diagonal case, the eigenvectors of $\mathcal{E}_{\mathrm{Cov}[x_k]}$ and $\mathcal{E}_{\mathrm{Cov}[\mathcal{C}(x_k)]}$ are slightly[8] different (as $\sqrt{\mathrm{Diag}\,(M)}$ and M are not

7. Note that the behavior for quantization, apart from the upper bound $\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M)$ is not quantified, it is thus possible that quantization performs even better than PP.
8. On the figure, there are nearly aligned, but actually differ.

jointly diagonalizable, as well as if $\mathrm{Diag}\,(M)$ is constant, although this case is not presented here, but in Figure S13 in Appendix E.3). They are equal for the diagonal case (as $\sqrt{\mathrm{Diag}\,(M)}$ and M are both diagonal so the eigenvectors are aligned with the axis). In both cases, the eigenvalue decay is reduced (from $\lambda_2/\lambda_1 = 1/10$ without compression to $1/\sqrt{10}$ with compression, which visually corresponds to a "wider" ellipse).

This slower eigenvalue decay results from the *unstructured-noise*, i.e., large noise on the weak-curvature direction, which is particularly visible on Figure 4. This is critical as it results in a potentially much larger limit rate, as $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_q, p_M)M^{-1}\right) \simeq \mathrm{Tr}\left(M^{-1/2}\right)$.

[Skt] For sketching, the eigenvectors remain the same for $\mathcal{E}_{\mathrm{Cov}[x_k]}$ and $\mathcal{E}_{\mathrm{Cov}[\mathcal{C}(x_k)]}$ (as I_2 and M are jointly diagonalizable, see Corollary 22), both in the diagonal and non-diagonal case. However, the isotropic noise with covariance I_2 is visible (wide ellipse), also drastically impacting $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M)M^{-1}\right) \propto \mathrm{Tr}\left(M^{-1}\right)$.

[Sp] For p-sparsification, eigenvectors are not aligned with the ones of M in the non-diagonal case, but are in the diagonal case. In this latter case, the covariance $\mathfrak{C}(\mathcal{C}_s, p_M)$ is proportional to M.

[Rd] Same remarks hold for Rand-1 than for sparsification. We see that $\mathfrak{C}(\mathcal{C}_{\mathrm{rd1}}, p_M)$ is diagonal, as expected. Again, both operators induce an unstructured-noise in the non-diagonal case.

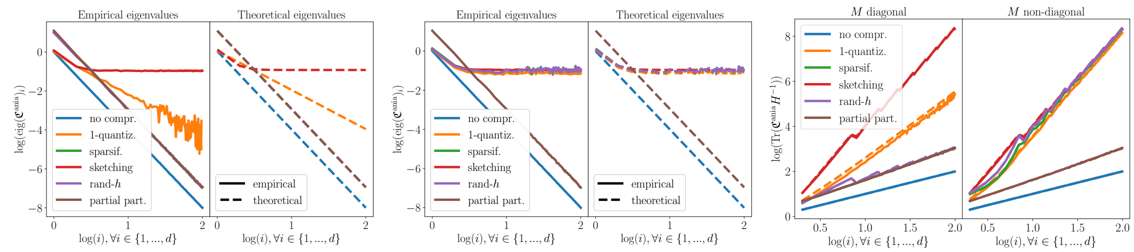
[PP] For PP, the covariances are always proportional (with factor p^{-1}), i.e., the ellipses have the same axis and $\mathcal{E}_{\mathrm{Cov}[\mathcal{C}(x_k)]}$ is a scaled version of $\mathcal{E}_{\mathrm{Cov}[x_k]}$.

We highlight the following points regarding pairwise comparisons:

- In the diagonal case, as stated by Item 1 in Proposition 24, $\mathrm{Cov}\left[\mathcal{C}_{\mathrm{s}}(x_k)\right]$ and $\mathrm{Cov}\left[\mathcal{C}_{\mathrm{PP}}(x_k)\right]$ are identical. $\mathrm{Cov}\left[\mathcal{C}_{\mathrm{rd1}}(x_k)\right]$ would have been identical too if $p = 1/d$ (but here we observe $\mathfrak{C}(\mathcal{C}_{\mathrm{rd1}}, p_M) \preccurlyeq \mathfrak{C}(\mathcal{C}_{s/\mathrm{PP}}, p_M)$ because the variance of rand-1 is smaller that for sparsification/PP).
- In the non-diagonal case, from Item 2 in Proposition 24, we have $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M)M^{-1}\right) \leq \mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_s, p_M)M^{-1}\right)$, however we do not have $\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) \preccurlyeq \mathfrak{C}(\mathcal{C}_s, p_M)$, hence we can not conclude anything on $\mathrm{Cov}\left[\mathcal{C}_{\mathrm{PP}}(x_k)\right]$ and $\mathrm{Cov}\left[\mathcal{C}_{\mathrm{s}}(x_k)\right]$.
- In the non-diagonal scenario, we observe on Figure 3, that $\mathfrak{C}(\mathcal{C}_q, p_M) \preccurlyeq \mathfrak{C}(\mathcal{C}_s, p_M)$ (as in Item 2 in Proposition 25).

3.3.2 Illustration of Proposition 21 and Corollary 22 in dimension $d > 2$

Another way of visualizing the structured and isotropic parts of the noise is by plotting the eigenvalues of $\mathfrak{C}(\mathcal{C}, p_M)$ in dimension $d = 100$. This is done in Figure 5, in which we plot the eigenvalues in decreasing order for both M and $\mathfrak{C}(\mathcal{C}, p_M)$, with Gaussian $p_M = \mathcal{N}(0, M)$ and $\mathrm{Sp}(M) = \{(1/i^4)_{i=1}^d\}$. We see that in the diagonal case, in Figure 5a, all operators but $\mathcal{C}_{\mathrm{q}}, \mathcal{C}_\Phi$ have a covariance proportional to M (thus a slope -4 on a log/log scale), while $\mathcal{C}_{\mathrm{q}}$ is proportional to $\sqrt{M}$ (thus a slope -2) and $\mathcal{C}_\Phi$ has an isotropic component (thus eigenvalues not decreasing to 0). In Figure 5b we see that only $\mathcal{C}_{\mathrm{PP}}$ has a covariance proportional to M while all other ones have an isotropic component (thus eigenvalues not decreasing to 0). We plot both empirical values and the ones obtained in Proposition 21, which shows that



(a) M diagonal, $d = 100$ (b) M non-diagonal, $d = 100$ (c) $\mathrm{Tr}\left(\mathfrak{C}_{\text{ania}} M^{-1}\right)$, $d \in [\![2, 100]\!]$

Figure 5: Figures 5a & 5b: Eigenvalues of $\mathfrak{C}(\mathcal{C}, p_M)$. Figure 5c: $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}, p_M) M^{-1}\right)$. $K = 10^4, \omega = 10$, $M = Q\mathrm{Diag}\left((1/i^4)_{i=1}^d\right) Q^T$ and $Q = \mathrm{I}_d$ (on 5a & 5c-l) or $Q \sim \mathrm{Unif}(\mathcal{O}_d)$ (on 5b & 5c-r). Plain lines: empirical values; dashed lines: theoretical formula or upper bound given by Proposition 21.

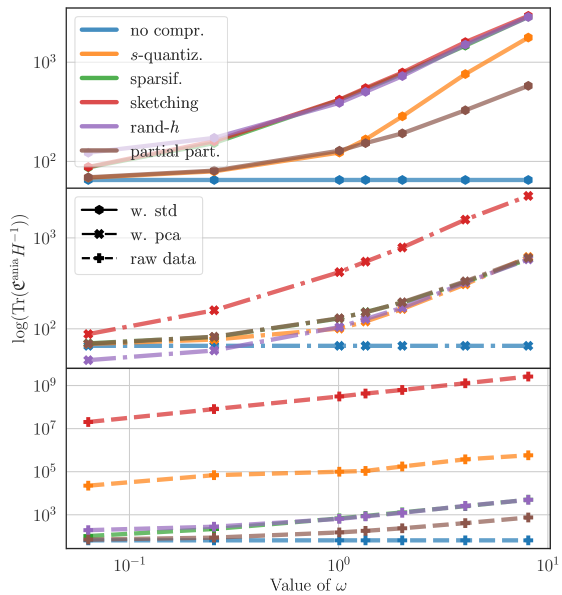
the upper bound on quantization is reasonable in practice and acts as a safety check for other compression schemes.

We plot on Figure 5c the theoretical and empirical $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}, p_M) M^{-1}\right)$ again in two cases, diagonal and non-diagonal. In the diagonal case, PP, sparsification, and rand-h have the same behavior; their traces have the smallest value among all compressors. However, in the general case of non-diagonal features' covariance, all compression operators have similar slow performance except for PP. For $d = 100$, all the compressors *have* $\omega = 10$, *but* $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}, p_M) M^{-1}\right)$ *varies by several orders depending on the compressor*, illustrating again that compressors satisfying Lemma 16 with the same ω may have vastly different behaviors.

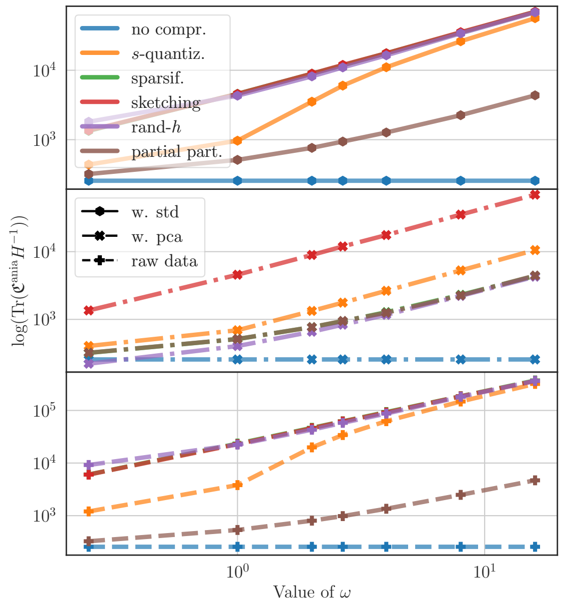
Lastly, we perform the same experiments on $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}, p_M) M^{-1}\right)$, but on non-simulated datasets, namely `quantum` (Caruana et al., 2004) and `cifar-10` (Krizhevsky et al., 2009): in Figure 6 we plot $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}, p_M) M^{-1}\right)$ w.r.t. the worst-case-variance-level ω of the compression in three scenarios: **(top-row)**—with data standardization, thus $\mathrm{Diag}\left(M\right)$ is constant equal to 1; **(middle-row)**—with a PCA, thus with a diagonal covariance M (note that this is for illustration purpose: performing a PCA would be more expensive computationally than running Algorithm 2); and **(bottom-row)**—without any data transformation. As a pre-processing, we have resized images of the `cifar-10` dataset to a 16×16 dimension. We adjust $h \in \mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_\Phi$, and $p \in \mathcal{C}_{\mathrm{PP}}, \mathcal{C}_{\mathrm{s}}$ to make ω vary. Besides, in order to also adjust ω for quantization, we use the s-quantization (Definition 26) schema which generalizes 1-quantization.

Definition 26 (s-quantization operator) *Given $z \in \mathbb{R}^d$, the s-quantization operator $\mathcal{C}_s$ is defined by $\mathcal{C}_s(z) := \mathrm{sign}(z) \times \|z\|_2 \times \frac{\chi}{s}$. $\chi \in \mathbb{R}^d$ is a random vector with j-th element defined as:* $\chi := \begin{cases} l+1 & \text{with probability } s\frac{|z_j|}{\|z\|_2} - l\,, \\ l & \text{otherwise} \end{cases}$ *where the level l is such that* $\dfrac{s|z_j|}{\|z\|_2} \in [l, l+1[$.

The s-quantization scheme verifies Assumption L.1 with $\omega = \min(d/s^2, \sqrt{d}/s)$. Proof can be found in (Alistarh et al., 2017, see Appendix A.1). We do not compute $\mathcal{M}_1, \mathcal{M}_2$ and the covariance $\mathfrak{C}_{\text{ania}}$.



(a) quantum (tabular dataset)



(b) cifar10 (images)

Figure 6: $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}, p_M) M^{-1}\right)$ w.r.t the level of ω for quantum and cifar10. X/Y-axis are in log scale. Note that the plots may have different magnitudes.

Interpretation. (Top-row): with standardization, the order predicted from Proposition 24.2 (large ω), and Proposition 25.2 (low ω) is obtained for both `quantum` and `cifar-10`: $\mathcal{C}_{\mathrm{PP}} \leq \mathcal{C}_{\mathrm{q}} \leq \mathcal{C}_{\mathrm{s}} \simeq \mathcal{C}_{\mathrm{rd}h} \simeq \mathcal{C}_{\Phi}$. For quantization, we observe two regimes: 1) when ω tends to zero, quantization and PP outperform sketching, sparsification, and rand-h, that are equivalent. 2) when ω increases, quantization changes from scaling as PP to scaling as the second group. **(Middle-row):** in the diagonal regime, comments made for Figure 5c-l are still valid. **(Bottom-row):** We observe that for a generic matrix M (obtained from raw-data) there is no systematic order between compression schemes. This is un-avoidable as the order for a "M diagonal" and "M with constant-diagonal" is *not* the same. We observe that:

- for `quantum`, $\quad \mathcal{C}_{\mathrm{PP}} \leq \mathcal{C}_{\mathrm{s}} \overset{\sim}{\leq} \mathcal{C}_{\mathrm{rd}h} \ll \mathcal{C}_{\mathrm{q}} \ll \mathcal{C}_{\Phi}$
- for `cifar-10`, $\mathcal{C}_{\mathrm{PP}} \ll \mathcal{C}_{\mathrm{q}} \ll \mathcal{C}_{\mathrm{s}} \simeq \mathcal{C}_{\mathrm{rd}h} \simeq \mathcal{C}_{\Phi}$.

We also observe that $\mathcal{C}_{\Phi}$, which is the only operator to always induce an isotropic component, may be much worse than all other compressors (e.g., on `quantum`). Ultimately, the order depends on the covariance matrix M. Here we observe that the raw-data behavior is close for `cifar-10` to the standardized version, while for `quantum` the order between compressors is the same for raw-data and diagonal (although the ratios are different). In Appendix E.4 (Table S3), we provide an illustration of the covariance matrices, that supports such interpretation.

3.4 Numerical experiments on Algorithm 2

In this section, we run Algorithm 2 on both synthetic and real datasets to illustrate the combined theoretical results of Sections 2 and 3. In Figure 7, we compare the compression operators to the baseline of no-compression. We plot the excess loss of the Polyak-Ruppert iterate $F(\overline{w}_k) - F(w)$, versus the index in log/log scale. Each experiment is conducted 5 times, with a new dataset generated from a new seed. The standard deviation of $\log_{10}(F(\overline{w}_k) - F(w))$ is indicated by the shadow-area.

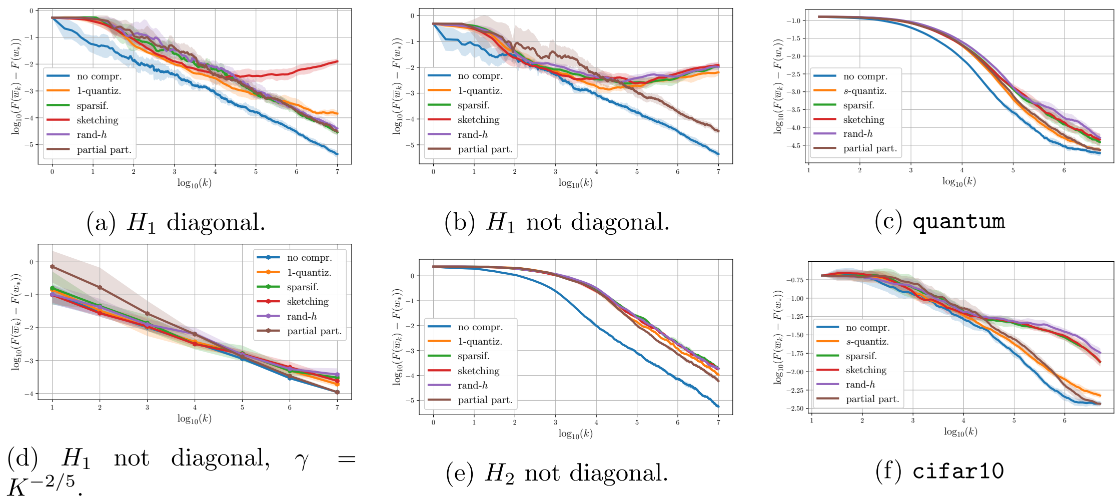
Setting: (a) *Synthetic dataset generation:* The dataset is generated using Model 2 with $K = 10^7$, $\sigma^2 = 1$, an optimal point w_* set as a constant vector of ones and a geometric eigenvalues decay of $D_1 = \mathrm{Diag}\left((1/i^4)_{i=1}^d\right)$ (resp. $D_2 = \mathrm{Diag}\left((1/i)_{i=1}^d\right)$). For $i \in \{1, 2\}$, the covariance matrix is $H_{\{i\}} = Q D_{\{i\}} Q^T$, where Q is either orthogonal matrix, or $Q = \mathrm{I}_d$ in the case of a diagonal features' matrix. (b) *Real datasets processing:* We resize images of the `cifar-10` dataset to a 16×16 dimension, and then for both datasets, we apply standardization. To compute the optimal point (and so to compute the excess loss), we run SGD over 200 passes on the whole dataset and consider the last Polyak-Ruppert average as the optimal point w_*. (c) *Algorithm 2:* We take a constant step-size $\gamma = 1/(2(\omega + 1)R^2)$ with R^2 the trace of the features' covariance, and $w_0 = 0$ as initial point. We set the batch-size $b = 1$ and the compressor variance $\omega = 10$ for synthetic datasets. For `cifar-10` and `quantum`, we run Algorithm 2 for 5×10^6 iterations (it corresponds to 100 passes on the whole dataset) with a batch-size $b = 16$, and using a s-quantization (Definition 26). We set $s = 16$ for `cifar-10` (factor 2 compression) and $s = 8$ for `quantum` (factor 4 compression), the compressor variance is therefore $\omega \approx 1$ for both datasets. These settings are summarized in Tables S1 and S2 in Appendix A.1. Additionally, to illustrate Corollary 9, we plot on Figure 7d the final excess loss after running Algorithm 2 with an horizon-dependent step-size $\gamma = K^{-2/5}$, computed for seven values of $K \in \{10^i, i \in [\![1, 7]\!]\}$.

Interpretation—H diagonal (Figure 7a). For sparsification, rand-h, and PP (linear compressors), the rate of convergence is given by Theorem 12. As stated by Corollary 22, the covariance $\mathfrak{C}_{\mathrm{ania}}$ is proportional to H leading to a $O(1/K)$ rate. We indeed observe in Figure 7a that excess loss is linear in a log/log scale.

For non-linear compression operators, the rate is given by Theorem 8. On the one hand, 1-quantization results in a slower eigenvalues' decay, leading to a larger $\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}} H^{-1}\right)$, thus a slower convergence than linear compressors. On the other hand, for sketching, covariance has a purely isotropic part scaling with I_d, which causes $\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}} H^{-1}\right)$ to strongly depend on the strong-convexity coefficient μ resulting in an extremely large constant. Both behaviors are observed in Figure 7a.

Interpretation—H not diagonal (Figures 7b and 7e). In the case of the high eigenvalues' decay of H_1 ($\mu = 10^{-8}$), the only compressor that shows in Figure 7b a linear rate of convergence in the log/log scale is PP. All others exhibit a saturation phenomenon after a certain number of iterations. This is again due to the unstructured part of the noise for all other compressors, as given by Proposition 21. Besides, we also note an increase of the excess loss after some iterations that is likely caused by the accumulation of noise on axis onto which the curvature of H is weak (but the isotropic noise is not). However, taking the optimal horizon-dependent step-size given by Corollary 9, we recover on Figure 7d for all compressor $\mathcal{C}$ the sub-linear convergence rate of PP shown at Figure 7b, reducing by

Synthetic dataset | Real datasets



(a) H_1 diagonal. (b) H_1 not diagonal. (c) `quantum`

(d) H_1 not diagonal, $\gamma = K^{-2/5}$. (e) H_2 not diagonal. (f) `cifar10`

Figure 7: Logarithm excess loss of the Polyak-Ruppert iterate for a single client ($N = 1$).

a factor 100 the excess loss w.r.t. to the scenario where $\gamma = 1/2(\omega + 1)R^2)$. While using a small step-size is slightly worse for SGD, it reduces the second and third term of the variance in Theorem 8 that depends on μ for other compressors. And in the scenario of a slow eigenvalues' decay ($\mu = 10^{-2}$), we observe on Figure 7e that all compressors reach the sub-linear rate (same slope -1 on the log/log plot), but with different constants. This illustrates Theorems 8 and 12 in the case of moderate coefficient μ where we expect the second and third parts of the variance term to be negligible.

Interpretation - real datasets, H with constant diagonal (Figures 7c and 7f). As we use s-quantization, this experience is going beyond Propositions 21 and 25 which only apply to 1-quantization. In the case of covariance with constant diagonal, Proposition 25 states that 1-quantization is better than projection-based compressors and comparable to partial participation. In practice, we observe that s-quantization performs competitively with PP and outperforms all other compressors. Besides, the asymptotic behavior is consistent with Figure 6 (top-row) for $\omega = 1$, where the order $\mathcal{C}_{\mathrm{PP}} \simeq \mathcal{C}_{\mathrm{q}} \ll \mathcal{C}_{\mathrm{s}} \simeq \mathcal{C}_{\mathrm{rd}h} \simeq \mathcal{C}_{\Phi}$ is observed.

3.5 Conclusion

In this section, we investigated how the compression scheme choice impacts the convergence rate, first by showing that quantization-based and projection-based methods respectively satisfy Theorem 8 and Theorem 12, resulting in different non-asymptotic behaviors. In the asymptotic regime, in both cases, the averaged excess loss scales as $\mathrm{Tr}\left(H^{-1}\mathfrak{C}_{\mathrm{ania}}\right)/K$. We then analyzed the impact of the most-used schemes on this limit rate. Overall, it appears that all compression schemes typically generate an *unstructured-noise*, which covariance does not scale with H, contrarily to the classical un-compressed Algorithm 1. The one exception is PP, which corresponds (on a single worker) to performing fewer iterations.

For other compression schemes, we show the impact of the covariance H: depending on the correlation between features (H diagonal or not) and on the pre-processing (e.g., standardization for which H has diagonal constant), the ordering between compression scheme varies. In many cases, this highlights the need for an additional regularisation when running Algorithm 2: all compression schemes (but PP) result in a significant noise that accumulates along the low curvature directions. Our results can be extended to the ridge (a.k.a., Tikhonov) regularized case (see Dieuleveut et al., 2017), which creates an additional bias but changes the rate $\mathrm{Tr}\left(H^{-1}\mathfrak{C}_{\mathrm{ania}}\right)/K$ into $\mathrm{Tr}\left((H+\lambda \mathrm{I})^{-1}\mathfrak{C}_{\mathrm{ania}}\right)/K$. The theoretical optimal choice for λ depending on H and the compression scheme could be obtained from our analysis but is left as future work.

We now turn to the distributed/federated case, which motivates the study of compression schemes for practical applications.

4. Application to Federated Learning

In this section, we consider Algorithm 3 under Model 1, which corresponds to heterogeneous Federated Learning on a network composed of N clients. We hereafter consider two particular cases naturally raising from Model 1: covariate-shift and optimal-point-shift. Note this results can easily be extended to the case of a heterogeneous level of noise by clients.

First, in Subsection 4.1, the *covariate-shift* case, i.e., Model 1 with $w_*^i = w_*$ for all i (thus the distribution of y^i conditional to x^i does not change between workers), but on the other hand, the features' marginal distributions are different, in particular, $H_i \neq H_j$. Second, in Subsection 4.2, the *optimal-point-shift* case, i.e., for each client $i, j \in [N]$, their optimal points are different $w_*^i \neq w_*^j$, but $H_i = H_j$. In the rest of the section, we denote $\overline{H} := \frac{1}{N}\sum_{i=1}^N H_i$, $\overline{R}^2 := \frac{1}{N}\sum_{i=1}^N R_i^2$, and we have $F(w_k) - F(w_*) = \frac{1}{2}\left\langle \eta_{k-1}, \overline{H}\eta_{k-1}\right\rangle$.

4.1 Heterogeneous covariance

In this section, we first show that Theorems 8 and 12 on (LSA) from Section 2 can be applied to the Federated Learning case within the scenario of covariate-shift. Corollary 27 is proved in Appendix F.1.

Corollary 27 (Algorithm 3 with covariate-shift) *Consider Algorithm 3 under Model 1 with $w_*^i = w_*^j$ (and potentially $H_i \neq H_j$).*

1. *For a compressor $\mathcal{C} \in \{\mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{sq}}, \mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\Phi}, \mathcal{C}_{\mathrm{PP}}\}$, Theorem 8 holds, with $H_F = \overline{H}$, $R_F^2 = \overline{R}^2$, $\mathcal{A} = (\omega+1)\overline{R}^2\sigma^2/N$, $\mathcal{M}_2 = (\omega+1)\max_{i\in[N]}(R_i^2)/N$, $\mathcal{M}_1 = \Omega\sigma\max_{i\in[N]}(R_i^2)/N$.*
2. *Moreover for any linear compressor $\mathcal{C} \in \{\mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\Phi}, \mathcal{C}_{\mathrm{PP}}\}$, Theorem 12 holds, with the same constants and $\text{Ш}_{\mathrm{add}} = \sigma^2\max_{i\in[N]}(\text{Ш}_{H_i})/N$ and $\text{Ш}_{\mathrm{mult}} = \max_{i\in[N]}(R_i^2\text{Ш}_{H_i})/N$, with $(\text{Ш}_{H_i})_{i=1}^N$ given in Corollary 18.*

The Hessian of the objective function is now $\overline{H}$, and Theorems 8 and 12 still hold. The proof consists in showing that with Lemma 16, Assumptions 1 to 4 on the resulting random field $(\xi_k)_{k\in\mathbb{N}^*}$ are valid, with the constants given above.

In order to understand the impact of the compressor on the limit convergence rate, we establish a formula for $\mathfrak{C}_{\mathrm{ania}}$ similar to Equation (4). In the setting of covariate-shift, we have for any clients $i, j \in [N]$, $w_*^i = w_*^j$, thus

$$\xi_k^{\text{add}} \overset{\text{def. 4}}{=} \xi_k(0) \overset{\text{algo 3}}{=} \nabla F(w_*) - \frac{1}{N} \sum_{i=1}^N \mathcal{C}_k^i(g_k^i(w_*))$$
$$\overset{\text{eq. 2}}{=} -\frac{1}{N} \sum_{i=1}^N \mathcal{C}_k^i((\langle x_k^i, w_* \rangle - y_k^i) x_k^i) \underset{\text{with } w_*^i = w_*^j}{\overset{\text{model 1}}{=}} \frac{1}{N} \sum_{i=1}^N \mathcal{C}_k^i(\varepsilon_k^i x_k^i)\,.$$

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:

$$\mathfrak{C}_{\text{ania}} = \mathbb{E}\left[(\xi_k^{\text{add}})^{\otimes 2}\right] = \mathbb{E}\left[\left(\frac{1}{N} \sum_{i=1}^N \mathcal{C}_k^i(\varepsilon_k^i x_k^i)\right)^{\otimes 2}\right] \overset{\text{indep. of } (\mathcal{C}_k^i)_{i=1}^d}{=} \frac{1}{N^2} \sum_{i=1}^N \mathbb{E}\left[\mathcal{C}_k^i(\varepsilon_k^i x_k^i)^{\otimes 2}\right]$$
$$= \frac{\sigma^2}{N^2} \sum_{i=1}^N \mathbb{E}\left[\mathcal{C}_k^i(x_k^i)^{\otimes 2}\right] \overset{\text{Def. 19}}{=} \frac{\sigma^2}{N^2} \sum_{i=1}^N \mathfrak{C}(\mathcal{C}_k^i, p_{H_i}) \overset{\text{notation}}{=:} \frac{\sigma^2}{N} \overline{\mathfrak{C}((\mathcal{C}^i, p_{H_i})_{i=1}^N)}\,. \tag{5}$$

The operator $\overline{\mathfrak{C}((\mathcal{C}^i, p_{H_i})_{i=1}^N)}$ generalizes the notion of *compressor's covariance* (Definition 19) to the case of multiple clients, and Equation (5) corresponds to Equation (4).

Remark 28 (All clients use the same *linear* compressor) *If for all $i \in [N]$, $\mathcal{C}^i \overset{(d)}{=} \mathcal{C}$ and $\mathcal{C} \in \{\mathcal{C}_{\text{PP}}, \mathcal{C}_{\text{s}}, \mathcal{C}_{\text{rd}h}, \mathcal{C}_\Phi\}$, leveraging Remark 23, we have*

$$\overline{\mathfrak{C}((\mathcal{C}^i, p_{H_i})_{i=1}^N)} = \mathfrak{C}(\mathcal{C}, \overline{H})\,.$$

The analysis of (LSA) *on a single worker made in Section 3 is still valid in this setting with now the Hessian of the problem being equal to the average of covariance $\overline{H}$. Corollary 27 and Equation* (5) *prove that the case of covariate-shift is identical to the centralized setting with a variance reduced by a factor N.*

Remark 29 (Varying compressor/compression-level, or non-linear compression) *In most other cases, the computation of $\frac{\sigma^2}{N}\overline{\mathfrak{C}((\mathcal{C}^i, p_{H_i})_{i=1}^N)} = \frac{\sigma^2}{N^2}\sum_{i=1}^N \mathfrak{C}(\mathcal{C}_i^i, p_{H_i})$ is possible using the results of Subsection 3.3*

Overall, in the covariate-shift case, most insights from the centralized case remain valid, especially, client sampling (i.e., PP) is the safest way to limit the impact of compression. Moreover, the trade-offs and ordering between compressors remain preserved, especially regimes in which quantization outperforms other competitors.

4.2 Heterogeneous optimal point

Hereafter, we focus on the case of heterogeneous optimal points and consider that all clients share the same covariance matrix, i.e. for any $i, j \in [N]$, $H_i = H$, but potentially $w_*^i \neq w_*^j$. This can be seen as a case of *concept-shift* (Kairouz et al., 2019), and we also refer to the situation as *optimal-point-shift*. This setting could eventually be combined with the covariate-shift case. Similarly, Theorems 8 and 12 on (LSA) from Section 2 can be applied.

Corollary 30 (Algorithm 3 with concept-shift) *Consider Algorithm 3 under Model 1 with $H_i = H_j$ (and potentially $w_*^i \neq w_*^j$).*

1. *For a compressor $\mathcal{C} \in \{\mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{sq}}, \mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\Phi}, \mathcal{C}_{\mathrm{PP}}\}$, Theorem 8 holds, with $H_F = H$, $R_F^2 = R^2$, $\mathcal{A} = \frac{R^2(\omega+1)}{N}(\kappa \mathrm{Tr}\,(H \mathrm{Cov}\,[W_*]) + \sigma^2)$ with $W_* \sim \mathrm{Unif}(\{w_*^i, i \in [N]\})$, $\mathcal{M}_2 = (\omega+1)^2/N$, and $\mathcal{M}_1 = \Omega R^2 \sigma / N$.*

2. *Moreover for any linear compressor $\mathcal{C} \in \{\mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\Phi}, \mathcal{C}_{\mathrm{PP}}\}$, Theorem 12 holds, with the same constants and $\text{Ш}_{\mathrm{add}} = \sigma^2 \text{Ш}_H / N$ and $\text{Ш}_{\mathrm{mult}} = R^2 \text{Ш}_H / N$, with Ш_H given in Corollary 18.*

Corollary 30 can be proved reusing computation made for Corollary 27 and using below Proposition 31. We next aim at computing the additive noise covariance. We note $g_{k,*}^i = g_k^i(w_*)$ the local stochastic gradient evaluated at optimal point w_*. We have, in Model 1, for any $w \in \mathbb{R}^d$, $F_i(w) := \mathbb{E}(\langle x_k^i, w - w_*^i \rangle - x_k^i \varepsilon_k^i)^2$, thus $\nabla F(w) = \frac{1}{N}\sum_{i=1}^N H(w - w_*^i)$, and $w_* = \sum_{i=1}^N w_*^i / N$. The setting of Definition 2 is verified with $H_F = H$, and for any $w \in \mathbb{R}^d$, that the random field ξ_k can be computed as:

$$\xi_k(w - w_*) \overset{\text{Def. 2\&Alg.3}}{=} H_F(w - w_*) - \frac{1}{N}\sum_{i=1}^N \mathcal{C}^i(g_k^i(w)), \text{ thus } \xi_k^{\mathrm{add}} \overset{\text{Def. 4}}{=} -\frac{1}{N}\sum_{i=1}^N \mathcal{C}^i(g_{k,*}^i),$$

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 $k \in \mathbb{N}$:

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

where p_{Θ_i} is the distribution of $g_{k,*}^i$ (for any k). In the last inequality, we simply discarded the non-positive term $-H\sum_{i=1}^N (w_* - w_*^i)^{\otimes 2} H$. For linear compressors, by Proposition 21, $\mathfrak{C}_{\mathrm{ania}}$ is a linear function of $\frac{1}{N}\sum_{i=1}^N \Theta_i$—the averaged second-order moment of the local gradients $(g_{k,*}^i)_{i=1}^N$. In order to bound this quantity, following Dieuleveut et al. (2017), we make the following assumption.

Assumption 5 *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:*

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

For instance, it is verified for Gaussian vectors with $\kappa = 3$. By Cauchy-Schwarz inequality, it implies that $\mathbb{E}[\langle z, x_1^i \rangle^2 \ (x_1^i)^{\otimes 2}] \preccurlyeq \kappa \langle z, Hz \rangle H$ for all $z \in \mathbb{R}^d$. We obtain the following proposition.

Proposition 31 (Impact of client-heterogeneity) *Let W_* be a random variable uniformly distributed over $\{w_*^i, i \in [N]\}$, thus such that,* $\mathrm{Cov}\,[W_*] = \frac{1}{N} \sum_{i=1}^N (w_* - w_*^i)^{\otimes 2}$*, then:*

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

Proof We have:

$$\begin{aligned}\Theta_i &= \mathbb{E}[((x_k^i \otimes x_k^i)(w_* - w_*^i) + x_k^i \varepsilon_k^i)^{\otimes 2}] \overset{(\varepsilon_k^i) \perp (x_k^i)}{=} \mathbb{E}[(x_k^i \otimes x_k^i)(w_* - w_*^i)^{\otimes 2} (x_k^i \otimes x_k^i)] + \sigma^2 H \\ &\overset{\text{Ass. 5}}{\preccurlyeq} \kappa \left\langle w_* - w_*^i, H(w_* - w_*^i) \right\rangle H + \sigma^2 H = \kappa \mathrm{Tr}\left(H (w_* - w_*^i)^{\otimes 2} \right) H + \sigma^2 H \,.\end{aligned}$$

■

In words, we have the following two main observations.

Remark 32 (Structured noise before compression) *Before compression is possibly applied, the noise remains structured, i.e., with covariance proportional to H, in the case of concept-shift. As a consequence, the rate for un-compressed Equation* (LSA) *will remain independent of the smallest eigenvalue of H. This remark extends to the case where $\mathcal{C}_{\mathrm{PP}}$ is applied.*

Remark 33 (Heterogeneous vs homogeneous case.) *Compared to the homogeneous case, in which $\Theta_i = \sigma^2 H_i$ and $\mathfrak{C}_{\mathrm{ania}} = \frac{\sigma^2}{N} \overline{\mathfrak{C}((\mathcal{C}^i, p_{H_i})_{i=1}^N)}$, the averaged second-order moment increases from $\sigma^2 H$ to $(\kappa \mathrm{Tr}\left(H \mathrm{Cov}\,[W_*] \right) + \sigma^2) H$, showing the impact of the dispersion of the optimal points $(w_*^i)_{i=1}^N$. This corresponds to the typical variance increase in the compressed heterogeneous SGD case (Mishchenko et al., 2019; Philippenko and Dieuleveut, 2020).*

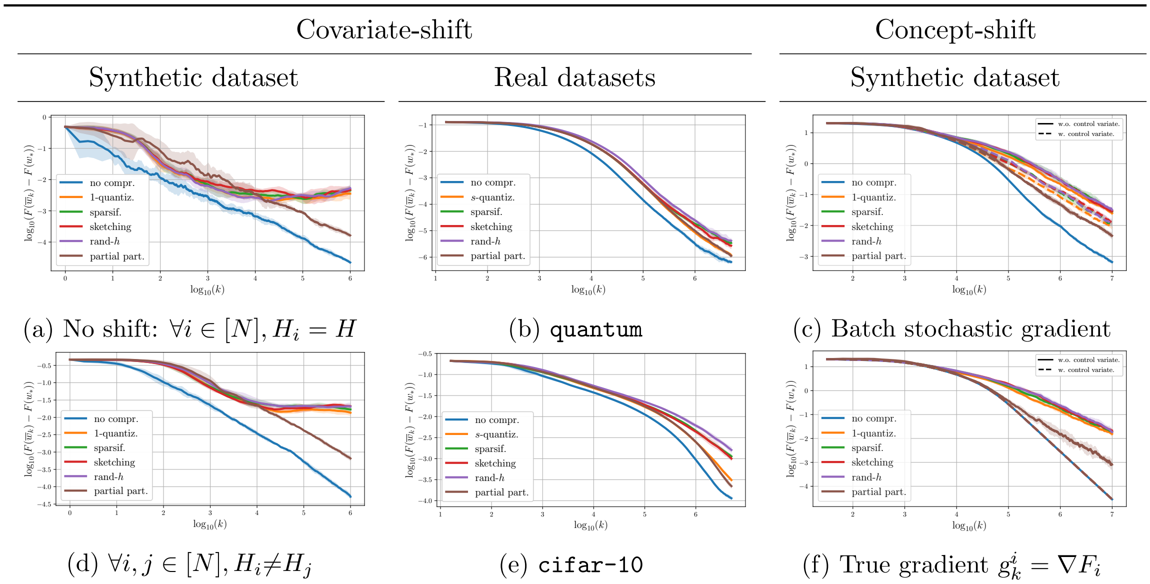
Concept-shift thus hinders the limit convergence rate. To limit this effect, a solution is to introduce a control-variate term $(h_k^i)_{k \in \mathbb{N}^*, i \in [N]}$, that is subtracted to the gradient before compression and asymptotically approximate $\nabla F_i(w_*)$ for any $i \in [N]$ (see Mishchenko et al., 2019). We explore this direction in Appendix F.2.

4.3 Numerical experiments

We support the theoretical results from Subsections 4.1 and 4.2 by performing experiments in the FL framework that extend the ones from Section 3.

On figures Figure 8, we present the results of the excess loss of the Polyak-Ruppert iterate $F(\overline{w}_k) - F(w_*)$ versus the number of iterations in log/log scale. The experiments were run 5 times, each time with different datasets (dispersion is shown by shaded area).

Settings. (a) *Synthetic dataset generation:* The dataset is generated using Model 1 with $N = 10$, $K = 10^6$ on each client, $\sigma^2/N = 1$. For any clients i in $[N]$, the covariance matrix is $H_i = Q_i D_i Q_i^T$, where Q_i is an orthogonal matrix. For heterogeneous clients, the dataset



(a) No shift: $\forall i \in [N], H_i = H$ (b) `quantum` (c) Batch stochastic gradient

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

Figure 8: Logarithm excess loss of the Polyak-Ruppert iterate iterations for $N = 10$ clients.

generation is as follows. *Covariate shift:* The rotation matrix Q_i is sampled independently for each client and the diagonal matrix D_i is Diag $\left((1/j^{\beta_i})_{j=1}^d\right)$ where $\beta_i \sim \mathrm{Unif}(\{3, 4, 5, 6\})$. *Concept-shift:* The optimal models of the clients $i \in [N]$ were drawn from a zero-centered normal distribution with a variance of $100\mathrm{I}_d$, that is, $w_*^i \sim \mathcal{N}(0, 100\mathrm{I}_d)$. We also take for all client i in $[N]$, $H_i = QDQ^T$, with $D = \mathrm{Diag}\left((1/j)\right)_{j=1}^d$. (b) *Real-dataset and covariate-shift:* To simulate non-i.i.d. clients, we split the dataset in heterogeneous groups (with equal number of points) using a K-nearest neighbors clustering on the TSNE representations (defined by Maaten and Hinton, 2008). Thus, the marginal feature distribution significantly varies between clients, providing a covariate-shift, while keeping the same distribution for the output conditional to the features on all clients. (c) *Algorithm 3:* We take a constant step-size $\gamma = 1/(2(\omega + 1)R^2)$ with $R^2 = \mathrm{Tr}\,(H)$ and $w_0 = 0$ as initial point. We set the batch-size $b = 1$ for synthetic datasets and $b = 16$ for real datasets, the compressor variance is $\omega = 10$. (d) *Algorithm 3* vs *Algorithm 4*: We take a bigger constant step-size $\gamma = (2R^2)^{-1}$ in order to emphasize the difference between the case w./w.o. control variate, we set $w_0 = 0$ as initial point and the compressor variance is $\omega = 10$. We set the batch-size $b = 32$ for Figure 8c and $b = K$ for Figure 8f.

Interpretation—homogeneous case and covariate-shift case (Figures 8a, 8b, 8d and 8e). These experiments extend those presented in Subsection 3.4 in the case of a single client. The observations made in the centralized case (Figure 7), especially on the impact of the compressor choice on the convergence and the ordering between limit convergence rates remain valid. This illustrates Corollary 27 and Remark 28: Theorems 8 and 12 hold in the case of homogeneous client or in the case of heterogeneous covariance and the only compressor that ensures that the noise is structured is client sampling (partial participation). On the real datasets, quantization is also competitive.

Interpretation—concept-shift case (Figures 8c and 8f). These experiments extend those presented on Figure 7e (slow eigenvalues' decay with $\mu = 10^{-2}$) to the scenario of concept shift. First, we observe on Figure 8c that for all compressors the convergence rate remains in $O(1/K)$, (though vanilla SGD converges faster during the first iterations). Second, we observe that control-variates improve convergence for compressors inducing unstructured noise ; this is predicted by theory, see Theorem S68. Third, on Figure 8f, at each iteration $k \in [K]$, we use deterministic gradients $g_k^i = \nabla F_i$ which leads to having a.s. $\xi_k^{\mathrm{add}} = 0$, and in the absence of compression, we obtain a $O(1/K^2)$ convergence rate for $\overline{w}_K$ which corresponds in Theorem 8 to the case where the dependency on the initial condition is dominated by $\frac{\|H_F^{-1/2}\eta_0\|^2}{\gamma^2 K^2}$. Overall, these experiments illustrate and support our theoretical insights.

5. Conclusion and open directions

Conclusion. In short, we investigate the impact of the choice of compression scheme on the convergence of the Polyak-Ruppert averaged iterate. By analysing the case of compressed least-squares regression, we shed light on the interplay between the Hessian of the optimization problem H_F, the features' distribution, the additive noise's covariance $\mathfrak{C}_{\mathrm{ania}}$, and the compression scheme. This shows fundamental differences between compression that deemed equivalent under the classical worst-case-variance assumption. We extend our analysis to the case of heterogeneous federated learning, a setting in which compression is widely used and its impact not fully understood.

More precisely, first, the analysis of the generic stochastic approximation algorithm (LSA) provides (1) the fact that projection based compressions achieve a faster convergence rate than quantization based, and that yet, their asymptotic rate is similar; (2) the analysis of quantization-based compression requires introducing a new Hölder-type regularity assumption for the analysis of the stochastic approximation scheme, and showing that such an assumption is satisfied for quantization.

Second, the computation of the additive noise's covariance $\mathfrak{C}_{\mathrm{ania}}$ reveals the impact of the compression scheme and the data distribution on the asymptotically dominant term. We obtain that (1) partial participation (i.e., client sampling in the federated case) is the only method that systematically ensures a convergence without a dependency on the strong-convexity constant; (2) other compressors may all induce an un-structured noise, with covariance scaling with I or $\sqrt{H}$, that strongly hinders convergence by accumulating noise on low curvature directions; (3) the relative performance or various schemes changes depending on the pre-processing applied to the data, making quantization the best method (apart from PP) when standardization is applied, but one of the worst (with random Gaussian projection) when the features are independent and the eigenvalues of the covariance decay rapidly (4) in that particular last setting, all projection based methods (but Gaussian projection) behave similarly.

Third, we discuss how these results apply to the federated case, that corresponds to the initial motivation. We show that we encompass two particular heterogeneity situations and how our analysis applies. Overall, these results are a step towards a better understanding of the impact of a widely used tool.

Open directions. This analysis can be extended to include various aspects that are beyond the scope of this work. First, one natural improvement for application in FL would be to consider also the scenario where each client runs several *local iterations* (McMahan et al., 2017; Karimireddy et al., 2020) before sending their updates, reducing further the cost of communication. Similar approach can be used, although the additive noise field would be more complicated, which potentially implies a different additive noise's covariance. Second, as mentioned in Subsection 3.5, our analysis could also be extended to the case of stochastic approximation with ridge regularization (e.g., following Dieuleveut et al., 2017) which in practice is helpful to mitigate the impact of the lack of strong convexity. Third, an obvious direction is to extend beyond quadratic functions and considering other objective functions, such as logistic regression or even shallow neural networks. Several results in the literature can be leveraged to tackle non quadratic but self-concordant losses Bach (2010); Gadat and Panloup (2023). Fourth, our analysis still only relies on second moments (variance and covariance) of the stochastic field. One major drawback of partial participation is to induce a significant increase on higher order moments. Incorporating higher order bounds may also bring novel insights to the use of compression in FL. Finally, all our analysis is made in finite dimension and our asymptotic focuses on $K \to \infty$: further works should analyze the case of infinite dimension: within the reproducing kernel Hilbert space (Dieuleveut and Bach, 2016) framework or within the overparametrized setting (Belkin et al., 2019).

6. Acknowledgements

We would like to thank Richard Vidal, Laeticia Kameni from Accenture Labs (Sophia Antipolis, France), and Eric Moulines from École Polytechnique for insightful discussions. We also thank Baptiste Goujaud for his great help in computing the covariance of sketching, and JM for her help on Figure 1. This research was supported by the *SCAI: Statistics and Computation for AI* ANR Chair of research and teaching in artificial intelligence, by *Hi!Paris*, and by *Accenture Labs* (Sophia Antipolis, France).

Constantin Philippenko thanks his sister Ania, whose covariance has greatly helped to write the theorems.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: a system for large-scale machine learning. In *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*, OSDI'16, pages 265–283, USA, November 2016. USENIX Association. ISBN 978-1-931971-33-1.

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

Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cedric Renggli. The Convergence of Sparsified Gradient Methods. *Advances in Neural*

Information Processing Systems, 31:5973–5983, 2018.

Reza Arablouei, Stefan Werner, Kutluyıl Doğançay, and Yih-Fang Huang. Analysis of a reduced-communication diffusion lms algorithm. *Signal Processing*, 117:355–361, 2015.

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4(none):384–414, January 2010. ISSN 1935-7524, 1935-7524. doi: 10.1214/09-EJS521. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

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.

Francis Bach. Lecture notes on statistical machine learning and convex optimization, 2022.

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.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. Publisher: National Acad Sciences.

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

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569. PMLR, 2018.

Neil Bershad. Analysis of the normalized lms algorithm with gaussian inputs. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4):793–806, 1986.

Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On Biased Compression for Distributed Learning. *arXiv:2002.12410 [cs, math, stat]*, February 2020. arXiv: 2002.12410.

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

Léon Bottou and Olivier Bousquet. The Tradeoffs of Large Scale Learning. *Advances in Neural Information Processing Systems*, 20:161–168, 2007.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A Benchmark for Federated Settings. *arXiv:1812.01097 [cs, stat]*, December 2019. arXiv: 1812.01097.

Luigi Carratino, Alessandro Rudi, and Lorenzo Rosasco. Learning with SGD and Random Features. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Rich Caruana, Thorsten Joachims, and Lars Backstrom. KDD-Cup 2004: results and analysis. *ACM SIGKDD Explorations Newsletter*, 6(2):95–108, December 2004. ISSN 1931-0145. doi: 10.1145/1046456.1046470.

Trishul Chilimbi, Yutaka Suzue, Johnson Apacible, and Karthik Kalyanaraman. Project adam: Building an efficient and scalable deep learning training system. In *11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14)*, pages 571–582, 2014.

Sélim Chraibi, Ahmed Khaled, Dmitry Kovalev, Peter Richtárik, Adil Salim, and Martin Takáč. Distributed fixed point methods with compressed iterates. *arXiv preprint arXiv:1912.09925*, 2019.

Xinyan Dai, Xiao Yan, Kaiwen Zhou, Han Yang, Kelvin KW Ng, James Cheng, and Yu Fan. Hyper-sphere quantization: Communication-efficient sgd for federated learning. *arXiv preprint arXiv:1911.04655*, 2019.

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.

Bernard Delyon. General results on the convergence of stochastic algorithms. *IEEE Transactions on Automatic Control*, 41(9):1245–1255, 1996.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *Ann. Statist.*, 44(4):1363–1399, 2016. ISSN 0090-5364. doi: 10.1214/15-AOS1391.

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.

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. doi: 10.1214/19-AOS1850. URL https://doi.org/10.1214/19-AOS1850.

Marie Duflo. *Random iterative models*, volume 34. Springer Science & Business Media, 1997.

Ilyas Fatkhullin, Igor Sokolov, Eduard Gorbunov, Zhize Li, and Peter Richtárik. EF21 with Bells & Whistles: Practical Algorithmic Extensions of Modern Error Feedback, October 2021. arXiv:2110.03294 [cs, math].

Nicolas Flammarion and Francis Bach. From averaging to acceleration, there is only a step-size. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 658–695, Paris, France, 03–06 Jul 2015. PMLR. URL https://proceedings.mlr.press/v40/Flammarion15.html.

Mark Iosifovich Freidlin and Alexander D Wentzell. Random perturbations. In *Random perturbations of dynamical systems*, pages 15–43. Springer, 1998.

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. ISSN 0304-4149. doi: 10.1016/j.spa.2022.11.012.

Venkata Gandikota, Daniel Kane, Raj Kumar Maity, and Arya Mazumdar. vqsgd: Vector quantized stochastic gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pages 2197–2205. PMLR, 2021.

Allen Gersho and Robert M Gray. *Vector quantization and signal compression*, volume 159. Springer Science & Business Media, 2012.

Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent. In *International Conference on Artificial Intelligence and Statistics*, pages 680–690. PMLR, 2020a.

Eduard Gorbunov, Dmitry Kovalev, Dmitry Makarenko, and Peter Richtarik. Linearly Converging Error Compensated SGD. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20889–20900. Curran Associates, Inc., 2020b.

László Györfi and Harro Walk. On the averaged stochastic approximation for linear regression. *SIAM Journal on Control and Optimization*, 34(1):31–61, 1996.

Farzin Haddadpour, Mohammad Mahdi Kamani, Aryan Mokhtari, and Mehrdad Mahdavi. Federated learning with compression: Unified analysis and sharp guarantees. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2350–2358. PMLR, 13–15 Apr 2021. URL `https://proceedings.mlr.press/v130/haddadpour21a.html`.

Ibrahim El Khalil Harrane, Rémi Flamary, and Cédric Richard. On reducing the communication cost of the diffusion lms algorithm. *IEEE Transactions on Signal and Information Processing over Networks*, 5(1):100–112, 2018.

Samuel Horvath, Chen-Yu Ho, Ludovit Horvath, Atal Narayan Sahu, Marco Canini, and Peter Richtárik. Natural compression for distributed deep learning. In *Mathematical and Scientific Machine Learning*, pages 129–141. PMLR, 2022.

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

Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random Design Analysis of Ridge Regression. In *Proceedings of the 25th Annual Conference on Learning Theory*, pages 9.1–9.24. JMLR Workshop and Conference Proceedings, June 2012. ISSN: 1938-7228.

Nikita Ivkin, Daniel Rothchild, Enayat Ullah, Vladimir Braverman, Ion Stoica, and Raman Arora. Communication-efficient Distributed SGD with Sketching. *Advances in Neural Information Processing Systems*, 32:13144–13154, 2019.

Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating Stochastic Gradient Descent for Least Squares Regression. In *Proceedings of the 31st Conference On Learning Theory*, pages 545–604. PMLR, July 2018a. ISSN: 2640-3498.

Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing Stochastic Gradient Descent for Least Squares Regression: Mini-batching, Averaging, and Model Misspecification. *Journal of Machine Learning Research*, 18(223): 1–42, 2018b. ISSN 1533-7928.

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

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error Feedback Fixes SignSGD and other Gradient Compression Schemes. In *International Conference on Machine Learning*, pages 3252–3261. PMLR, May 2019. ISSN: 2640-3498.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

Sarit Khirirat, Hamid Reza Feyzmahdavian, and Mikael Johansson. Distributed learning with compressed gradients. *arXiv preprint arXiv:1806.06573*, 2018.

Anastasia Koloskova, Sebastian Stich, and Martin Jaggi. Decentralized Stochastic Optimization and Gossip Algorithms with Compressed Communication. In *International Conference on Machine Learning*, pages 3478–3487. PMLR, May 2019. ISSN: 2640-3498.

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

Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtarik, Ananda Theertha Suresh, and Dave Bacon. Federated Learning: Strategies for Improving Communication Efficiency. In *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

Dmitry Kovalev, Elnur Gasanov, Alexander Gasnikov, and Peter Richtarik. Lower bounds and optimal algorithms for smooth and strongly convex decentralized optimization over time-varying networks. *Advances in Neural Information Processing Systems*, 34:22325–22335, 2021.

Alex Krizhevsky, Geoffrey Hinton, and others. Learning multiple layers of features from tiny images. 2009. Publisher: Citeseer.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Ensemble Nystrom Method. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

Harold J. Kushner and George Yin. Stochastic approximation and recursive algorithms and applications. 2003.

Louis Leconte, Aymeric Dieuleveut, Edouard Oyallon, Eric Moulines, and Gilles Pages. DoStoVoQ: Doubly Stochastic Voronoi Vector Quantization SGD for Federated Learning. May 2021.

Adrien Marie Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes.* Firmin Didot, Libraire pour les mathematiques, 1806.

Rémi Leluc and François Portier. Sgd with coordinate sampling: Theory and practice. *Journal of Machine Learning Research*, 23(342):1–47, 2022.

Ping Li, Trevor J Hastie, and Kenneth W Church. Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 287–296, 2006.

Shengqiao Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics & Statistics*, 4(1):66–70, 2010.

Zhize Li and Peter Richtárik. Canita: Faster rates for distributed convex optimization with communication compression. *Advances in Neural Information Processing Systems*, 34: 13770–13781, 2021.

Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtarik. Acceleration for Compressed Gradient Descent in Distributed and Federated Optimization. In *International Conference on Machine Learning*, pages 5895–5904. PMLR, November 2020. ISSN: 2640-3498.

Junhong Lin and Lorenzo Rosasco. Optimal Rates for Learning with Nyström Stochastic Gradient Methods. *arXiv preprint arXiv:1710.07797*, 2017.

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.

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

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

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. ISSN ISSN 1533-7928.

Odile Macchi. *Adaptative processing: the least mean squares approach with applications in transmission*, volume 71. New York: John Wiley & Sons, Ltd, 1995.

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.

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.

Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* 2018.

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

Nicole Muecke, Gergely Neu, and Lorenzo Rosasco. Beating SGD Saturation with Tail-Averaging and Minibatching. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Gergely Neu and Lorenzo Rosasco. Iterate averaging as regularization for stochastic gradient descent. In *Conference On Learning Theory*, pages 3222–3242. PMLR, 2018.

Julie Nutini, Mark Schmidt, Issam Laradji, Michael Friedlander, and Hoyt Koepke. Coordinate descent converges faster with the gauss-southwell rule than random selection. In *International Conference on Machine Learning*, pages 1632–1641. PMLR, 2015.

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.

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

Boris Polyak and Anatoli Juditsky. Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization*, 30:838–855, July 1992. doi: 10.1137/0330046.

Michael G Rabbat and Robert D Nowak. Quantized incremental algorithms for distributed optimization. *IEEE Journal on Selected Areas in Communications*, 23(4):798–808, 2005.

Ali Rahimi and Benjamin Recht. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

Ali Ramezani-Kebrya, Fartash Faghri, Ilya Markov, Vitalii Aksenov, Dan Alistarh, and Daniel M Roy. Nuqsgd: Provably communication-efficient data-parallel sgd via nonuniform quantization. *The Journal of Machine Learning Research*, 22(1):5074–5116, 2021.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. FedPAQ: A Communication-Efficient Federated Learning Method with Periodic Averaging and Quantization. In *International Conference on Artificial Intelligence and Statistics*, pages 2021–2031. PMLR, June 2020. ISSN: 2640-3498.

Peter Richtarik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A New, Simpler, Theoretically Better, and Practically Faster Error Feedback. In *Advances in Neural Information Processing Systems*, volume 34, pages 4384–4396. Curran Associates, Inc., 2021.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *Annals of Mathematical Statistics*, 22(3):400–407, September 1951. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177729586. Number: 3 Publisher: Institute of Mathematical Statistics.

Alessandro Rudi and Lorenzo Rosasco. Generalization Properties of Learning with Random Features. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is More: Nyström Computational Regularization. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. FALKON: An Optimal Large Scale Kernel Method. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and Communication-Efficient Federated Learning From Non-i.i.d. Data. *IEEE Transactions* *on Neural Networks and Learning Systems*, pages 1–14, 2019. ISSN 2162-2388. doi: 10.1109/TNNLS.2019.2944481. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

Frank Seide and Amit Agarwal. CNTK: Microsoft's Open-Source Deep-Learning Toolkit. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 2135, New York, NY, USA, August 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2945397.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Fifteenth Annual Conference of the International Speech Communication Association*. Citeseer, 2014.

Panagiotis Sidiropoulos. N-sphere chord length distribution. *arXiv preprint arXiv:1411.5639*, 2014.

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

Nikko Strom. Scalable distributed DNN training using commodity GPU cloud computing. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

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.

Aditya Varre and Nicolas Flammarion. Accelerated SGD for Non-Strongly-Convex Least Squares. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pages 2062–2126. PMLR, June 2022. ISSN: 2640-3498.

Santosh S Vempala. *The random projection method*, volume 65. American Mathematical Soc., 2005.

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

An Xu, Zhouyuan Huo, and Heng Huang. Optimal gradient quantization condition for communication-efficient distributed training. *arXiv preprint arXiv:2002.11082*, 2020.

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.

Supplementary material

In this appendix, we provide additional information to supplement our work. In Appendix A, we begin by detailing technical results, by introducing an auxiliary lemma and by proving Proposition 7 which gives a CLT for (LSA). Secondly, in respectively Appendix B and Appendix C, we give the proof of Theorems 8 and 12. Thirdly, in Appendix D, we verify that the setting of single-client compressed LSR fulfills the setting presented in Section 2. In Appendix E we prove that Lemma 16 hold and compute the compressors' covariance to establish Proposition 21 and Corollary 22. Finally, in Appendix F, we provide demonstrations for the federated learning case, including verifying assumptions (covariate-shift scenario) on random fields in Appendix F.1, and proving a Central Limit Theorem S68 in Appendix F.2 for the concept-shift scenario.

Additional notations. We use the Frobenius norm $\|A\|^2 := \mathrm{Tr}\left(A^\top A\right)$, which is the same notation as the vector Euclidean norm (no ambiguity in general), J_r to denote the $d \times d$ diagonal matrix whose r first diagonal elements are equal to one and all the other matrix's coefficients equal to zero, $\mathcal{S}_d^{++}(\mathbb{R})$ the cone of positive definite symmetric matrices, and $\mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ the set of random vectors defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ such that $\mathbb{E}[\|X\|^p] < \infty$. We define also the operator norm $|||A||| := \sqrt{\max \mathrm{eig}(A^\top A)}$.

Contents

A Technical results **37**

A.1 Settings of experiments . . . 37

A.2 Useful identities and inequalities . . . 37

A.3 An auxiliary inequality . . . 38

A.4 Asymptotic results: central limit theorem for (LSA) . . . 39

B Generalization of Bach and Moulines (2013). **41**

B.1 Proof principle . . . 41

B.2 Two bounds . . . 42

B.3 Final theorem . . . 46

C Generalisation of Bach and Moulines (2013) for linear multiplicative noise. **47**

C.1 Proof principle . . . 47

C.2 Lemmas for the noise process . . . 48

C.3 Final theorem . . . 53

D Validity of the assumptions made on the random fields **56**

E Compression operators **61**

E.1 Computation of the variance and covariance of the compression operators . 61

E.2 Variance and covariance of sketching . . . 64

E.3 Proof of Propositions 24 and 25 . . . 69

E.4 Empirical covariances computed on quantum and cifar10 . . . 72

F Technical results on federated learning. **73**

F.1 Validity of the assumptions made on the random fields in the case of covariate-shift . . . 73
F.2 Heterogeneous optimal point . . . 77

Appendix A. Technical results

A.1 Settings of experiments

In Tables S1 and S2, we summarize the settings of experiments presented in Subsection 3.4.

Table S1: Settings of experiments for a single client ($N = 1$) on synthetic data (Figures 7a and 7b).

Parameter	K	d	$\mathrm{eig}(H)_i$	w_*	σ^2	ω	γ^{-1}	w_0	$\#runs$
Values	10^7	100	$1/i^4$	$(1)_{i=1}^d$	1	10	$2(\omega+1)R^2$	0	5

Table S2: Settings of experiments for a single client ($N = 1$) on real data (Figures 7c and 7f).

Dataset	d	standardization	b	ω	γ^{-1}	w_0	$\#runs$	reference
`quantum`	65	✓	16	1	$2(\omega+1)R^2$	0	5	(CTL04)
`cifar-10`	256							(Kri09)

A.2 Useful identities and inequalities

In this Subsection, we recall some classical inequalities and results.

Inequality 1 *Let $N \in \mathbb{N}$ and $d \in \mathbb{N}$. For any sequence of vector $(a_i)_{i=1}^N \in \mathbb{R}^d$, we have the following inequalities:*

$$\left\| \sum_{i=1}^N a_i \right\|^2 \leq \left(\sum_{i=1}^N \|a_i\| \right)^2 \leq N \sum_{i=1}^N \|a_i\|^2 \,.$$

The first part of the inequality corresponds to the triangular inequality, while the second part is Cauchy's inequality.

Inequality 2 *Let x in $\mathbb{R}^d$ and A in $\mathcal{M}_{d,d}(\mathbb{R})$, then we have $\|Ax\| \leq \|\!|A|\!\|\|x\|$.*

Below, we recall Minkowski's and Jensen's inequalities. Additionally, we recall the Cauchy-Schwarz inequality for conditional expectations.
Let a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with Ω a sample space, $\mathcal{A}$ a σ-algebra, and $\mathbb{P}$ a probability measure.

Minkowski's inequality. Let $p > 1$ and suppose that X, Y are two random variables in $\mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$ (i.e. their p^{th} moment is bounded), we have the following triangular inequality:

$$\mathbb{E}[\|X+Y\|^p]^{1/p} \leq \mathbb{E}[\|X\|^p]^{1/p} + \mathbb{E}[\|Y\|^p]^{1/p} \,. \tag{S6}$$

Jensen's inequality. Suppose that $X : \Omega \longrightarrow \mathbb{R}^d$ is a random variable, then for any convex function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ we have:

$$f\left(\mathbb{E}(X)\right) \leq \mathbb{E} f(X) \,. \tag{S7}$$

Cauchy-Schwarz's inequality for conditional expectations. Suppose that X, Y are two random variables in $\mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$ (i.e. their second moment is bounded), then for any σ-algebra $\mathcal{F} \subset \mathcal{A}$ we have a.s.:

$$\mathbb{E}\left[XY \mid \mathcal{F}\right]^2 \leq \mathbb{E}\left[X^2 \mid \mathcal{F}\right] \mathbb{E}\left[Y^2 \mid \mathcal{F}\right] \,. \tag{S8}$$

Convergence in L^p-norm. Suppose that $(X_n)_{n\in\mathbb{N}}$ is a sequence of random variables in $\mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$, and that X is a random variable in $\mathcal{L}^p(\Omega, \mathcal{A}, \mathbb{P})$. We say that $(X_n)_{n\in\mathbb{N}}$ converges in L^p-norm towards X if $\mathbb{E}(\|X_n - X\|^p) \xrightarrow[n\to+\infty]{} 0$, it is denoted by: $X_n \xrightarrow[n\to+\infty]{L^p} X$.

In Subsection 3.3, we use ellipses to visual quadratic functions, therefore we provide in Definition S34 the mathematical definition.

Definition S34 (Representing positive matrices through ellipsoids) *Any symmetric positive definite matrix M in $\mathcal{S}_d^{++}(\mathbb{R})$ defines an ellipsoid $\mathcal{E}_M = \{x \in \mathbb{R}^d, x^\top M^{-1} x = 1\}$ centered around zero. The eigenvectors of M are the principal axes of the ellipsoid, and the squared root of the eigenvalues are the half-lengths of these axes. The ellipse corresponds to the sphere of radius 1 associated with the norm $N_{M^{-1}} = \sqrt{x^\top M^{-1} x}$.*

A.3 An auxiliary inequality

In this Section, we provide an auxiliary lemma that is specific to the framework considered in Section 2. It will be used in the proof of Theorem 8 and corresponds to an adaptation of Lemma 1 from Bach and Moulines (2013).

Lemma S35 (Auxiliary inequality on $\sum_{k=1}^K \mathbb{E}[\|H_F^{1/2}\eta_k\|^2]/K$) *Under Assumptions 1 and 2.1, for any K in $\mathbb{N}^*$ and any step-size $\gamma \in \mathbb{R}^+$ s.t. $\gamma(R_F^2 + 2\mathcal{M}_2) \leq 1$, the sequence $(w_k)_{k\in\mathbb{N}^*}$ produced by a setting such as in Definition 2, verifies the following bound:*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[\|H_F^{1/2}(w_k - w_*)\|^2] \leq \frac{\|\eta_0\|^2}{2\gamma K(1-\gamma(R_F^2 + 2\mathcal{M}_2))} + \frac{5\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)} \,.$$

Proof Let k in $\mathbb{N}^*$, we start writing that by Definition 2, we have $w_k = w_{k-1} - \gamma\nabla F(w_{k-1}) + \gamma\xi_k(\eta_{k-1})$. Thus taking the squared norm and developing it, gives:

$$\|\eta_k\|^2 = \|\eta_{k-1}\|^2 - 2\gamma\left\langle \eta_{k-1}, \nabla F(w_{k-1}) - \xi_k(\eta_{k-1})\right\rangle + \gamma^2 \left\|\nabla F(w_{k-1}) - \xi_k(\eta_{k-1})\right\|^2 \,. \tag{S9}$$

We need to bound the last term. By Definition 4, we have $\xi_k(\eta_{k-1}) = \xi_k^{\text{mult}}(\eta_{k-1}) + \xi_k^{\text{add}}$, hence using Inequality 1, we have:

$$\|\nabla F(w_{k-1}) - \xi_k(\eta_{k-1})\|^2 \leq 2\|\nabla F(w_{k-1}) - \xi_k^{\text{mult}}(\eta_{k-1})\|^2 + 2\|\xi_k^{\text{add}}\|^2 \,,$$

taking expectation w.r.t the σ-algebra $\mathcal{F}_{k-1}$, developping $\left\|\nabla F(w_{k-1}) - \xi_k^{\mathrm{mult}}(\eta_{k-1})\right\|^2$ and because $\mathbb{E}\left[\xi_k^{\mathrm{mult}}(\eta_{k-1}) \;\middle|\; \mathcal{F}_{k-1}\right] = 0$ (the random fields $(\xi_k)_{k\in\mathbb{N}^*}$ are zero-centered, see Definition 2), we have:

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla F(w_{k-1}) - \xi_k(\eta_{k-1})\|^2 \;\middle|\; \mathcal{F}_{k-1}\right] \\
&\qquad \leq 2\mathbb{E}\left[\|\nabla F(w_{k-1})\|^2 \mid \mathcal{F}_{k-1}\right] + 2\mathbb{E}\left[\|\xi_k^{\mathrm{mult}}(\eta_{k-1})\|^2 \;\middle|\; \mathcal{F}_{k-1}\right] + 2\mathbb{E}\left[\|\xi_k^{\mathrm{add}}\|^2 \;\middle|\; \mathcal{F}_{k-1}\right].
\end{aligned}
$$

Now, we use Definition 2 and Assumptions 1 and 2.1: it leads to:

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla F(w_{k-1}) - \xi_k(\eta_{k-1})\|^2 \;\middle|\; \mathcal{F}_{k-1}\right] \\
&\qquad \leq 2R_F^2\|H_F^{1/2}\eta_{k-1}\|^2 + 4\mathcal{M}_2\|H_F^{1/2}\eta_{k-1}\|^2 + 8\mathcal{A} + 2\mathcal{A} \\
&\qquad \leq 2(R_F^2 + 2\mathcal{M}_2)\|H_F^{1/2}\eta_{k-1}\|^2 + 10\mathcal{A}\,.
\end{aligned}
$$

Because the sequence of random field $(\xi_k)_{k\in\mathbb{N}^*}$ is zero-centered (Definition 2), we have:

$$
\mathbb{E}\left[\langle \eta_{k-1}, \nabla F(w_{k-1}) - \xi_k(\eta_{k-1})\rangle \mid \mathcal{F}_{k-1}\right] = \langle \eta_{k-1}, H_F\eta_{k-1}\rangle = \|H_F^{1/2}\eta_{k-1}\|^2\,,
$$

hence back to Equation (S9), we obtain:

$$
\mathbb{E}\left[\|\eta_k\|^2 \;\middle|\; \mathcal{F}_{k-1}\right] \leq \|\eta_{k-1}\|^2 - 2\gamma(1-\gamma(R_F^2 + 2\mathcal{M}_2))\|H_F^{1/2}\eta_{k-1}\|^2 + 10\mathcal{A}\gamma^2\,. \tag{S10}
$$

It follows that if $\gamma(R_F^2+2\mathcal{M}_2) \leq 1$, summing the previous bound and taking full expectation gives:

$$
\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|H_F^{1/2}\eta_{k-1}\|^2\right] \leq \frac{\|\eta_0\|^2 - \mathbb{E}\|\eta_K\|^2}{2\gamma K(1-\gamma(R_F^2 + 2\mathcal{M}_2))} + \frac{5\mathcal{A}\gamma}{1-\gamma(R_F^2 + 2\mathcal{M}_2)}\,,
$$

which allows concluding. ■

A.4 Asymptotic results: central limit theorem for (LSA)

The demonstration of Proposition 7 uses the following theorem from Polyak and Juditsky (1992) guaranteeing the asymptotic normality of the Polyak-Ruppert iterate.

Theorem S36 ***From Polyak and Juditsky (1992, see Theorem 1).***
For k in $\mathbb{N}^*$, we denote $\eta_k = w_k - w_*$ and we define $w_k = w_{k-1} - \gamma_k\nabla F(w_{k-1}) + \gamma_k\xi(\eta_{k-1})$. If we assume that:

- $\gamma_k \xrightarrow[k\to+\infty]{} 0$ *and* $\gamma_k^{-1}(\gamma_k - \gamma_{k+1}) = \underset{k\to+\infty}{o}(\gamma_k)$,
- *F is strongly convex and* $\left\|\nabla^2 F\right\|_\infty < \infty$,
- *the convergence in probability of the conditional covariance to a matrix Σ holds, i.e., we have a.s.* $\mathbb{E}[\xi(\eta_{k-1})\xi(\eta_{k-1})^\top \mid \mathcal{F}_{k-1}] \xrightarrow[k\to+\infty]{\mathbb{P}} \Sigma\,.$

Then for any K in $\mathbb{N}^*$, we have the asymptotic normality of $(\sqrt{K}\eta_{K-1})_{K\in\mathbb{N}^*}$*:*

$$\sqrt{K}\overline{\eta}_{K-1} \xrightarrow[K\to+\infty]{\mathcal{L}} \mathcal{N}(0,\Sigma^*) \text{ with } \Sigma^\star = \left\{\nabla^2 F\left(w_*\right)\right\}^{-1}\Sigma\left\{\nabla^2 F\left(w_*\right)\right\}^{-1}.$$

Below we present our CLT that gives the asymptotic normality of $(\sqrt{K}\eta_{K-1})_{K\in\mathbb{N}^*}$ in the case of strongly-convex case and decreasing step size.

Proposition S37 (CLT for (LSA)—strongly convex-case, deacreasing step-size) *Under Assumptions 1 and 2, consider a sequence $(w_k)_{k\in\mathbb{N}^*}$ produced in the setting of Definition 2 using a step-size $(\gamma_k)_{k\in\mathbb{N}^*}$ s.t. $\gamma_k = k^{-\alpha}$, $\alpha\in(0,1)$. Then $(\eta_K)_{K\geq 0}$ converges in L^2-norm to 0, i.e. $\eta_K \xrightarrow[K\to+\infty]{L^2} 0$.*
Furthermore, $(\sqrt{K}\overline{\eta}_{K-1})_{K\geq 0}$ is asymptotically normal with mean zero and covariance such that:

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

Proof
First, we have that in the case of decreasing step size s.t. for any k in $\mathbb{N}$, $\gamma_k = k^{-\alpha}$, we have: $\eta_K \xrightarrow[K\to+\infty]{L^2} 0$. This is a classical computation for SGD with bounded variance (Assumptions 1 and 2.1.). Detailed computations are for instance given in lectures notes of Bach (2022, pages 164-167 and 182), and Kushner and Yin (2003). To apply Theorem 1 from Polyak and Juditsky (1992, recalled in Theorem S36), which gives the desired result, it suffices to prove the convergence in probability of the covariance of the noise $\xi_k(\eta_{k-1})$ towards $\mathfrak{C}_{\text{ania}}$, as $k\to\infty$.

In the following, we show that $\lim_{k\to+\infty}\mathbb{E}\left[\xi_k(\eta_{k-1})\xi_k(\eta_{k-1})^\top \mid \mathcal{F}_{k-1}\right] \overset{\mathbb{P}}{=} \mathfrak{C}_{\text{ania}}$. We start writing:

$$\begin{aligned}\xi_k(\eta_{k-1})\xi_k(\eta_{k-1})^\top &= (\xi_k^{\text{add}} - \xi_k^{\text{mult}}(\eta_{k-1}))(\xi_k^{\text{add}} - \xi_k^{\text{mult}}(\eta_{k-1}))^\top \\ &= (\xi_k^{\text{add}})^{\otimes 2} - \xi_k^{\text{add}}\xi_k^{\text{mult}}(\eta_{k-1})^\top - \xi_k^{\text{mult}}(\eta_{k-1})(\xi_k^{\text{add}})^\top + \xi_k^{\text{mult}}(\eta_{k-1})^{\otimes 2}.\end{aligned}$$

(i) First, from Definition 4, it flows that $\mathbb{E}\left[\xi_k^{\text{add}}\otimes\xi_k^{\text{add}} \mid \mathcal{F}_{k-1}\right] = \mathfrak{C}_{\text{ania}}$.
(ii) Second, we show that $\mathbb{E}[\xi_k^{\text{mult}}(\eta_{k-1})^{\otimes 2} \mid \mathcal{F}_{k-1}]$ converges to 0 in probability: it is sufficient to show that: $\mathbb{E}[\|\xi_k^{\text{mult}}(\eta_{k-1})^{\otimes 2}\|_F \mid \mathcal{F}_{k-1}] \xrightarrow[k\to+\infty]{} 0$. To do so, we use the fact that $\|\xi_k^{\text{mult}}(\eta_{k-1})^{\otimes 2}\|_F = \|\xi_k^{\text{mult}}(\eta_{k-1})\|_2^2$, then with Assumption 2.2: $\mathbb{E}[\|\xi_k^{\text{mult}}(w-w_*)\|^2 \mid \mathcal{F}_{k-1}] \leq \mathcal{M}_1\|H^{1/2}\eta_{k-1}\| + \mathcal{M}_2\|H^{1/2}\eta_{k-1}\|^2$. And we have the result as we showed that $\eta_{k-1} \xrightarrow[k\to+\infty]{L^2} 0$.

(iii) Third, it remains to show that $\mathbb{E}[\xi_k^{\text{mult}}(\eta_{k-1})(\xi_k^{\text{add}})^\top \mid \mathcal{F}_{k-1}] \xrightarrow[k\to+\infty]{L^1} 0$. We use the Cauchy-Schwarz inequality's S8 for conditional expectation:

$$\begin{aligned}\mathbb{E}\left[\|\xi_k^{\text{mult}}(\eta_{k-1})(\xi_k^{\text{add}})^\top\|_F \;\middle|\; \mathcal{F}_{k-1}\right]^2 &= \mathbb{E}\left[\|\xi_k^{\text{mult}}(\eta_{k-1})\|_2\|(\xi_k^{\text{add}})^\top\|_2 \;\middle|\; \mathcal{F}_{k-1}\right]^2 \\ &\leq \mathbb{E}\left[\|\xi_k^{\text{mult}}(\eta_{k-1})\|_2^2 \;\middle|\; \mathcal{F}_{k-1}\right]\mathbb{E}\left[\|\xi_k^{\text{add}}\|_2 \;\middle|\; \mathcal{F}_{k-1}\right].\end{aligned}$$

The sequence of random vectors $(\xi_k^{\mathrm{add}})_{k\in\mathbb{N}^*}$ is i.i.d., and moreover we have shown previously that $\mathbb{E}[\|\xi_k^{\mathrm{mult}}(\eta_{k-1})\|^2 \mid \mathcal{F}_{k-1}]$ tends to 0, hence $\mathbb{E}[\xi_k^{\mathrm{mult}}(\eta_{k-1})(\xi_k^{\mathrm{add}})^\top \mid \mathcal{F}_{k-1}]$ converges to 0 in probability. Consequently, we can state that $\mathbb{E}[\xi_k(\eta_{k-1})^{\otimes 2} \mid \mathcal{F}_{k-1}] \xrightarrow[k\to+\infty]{\mathbb{P}} \mathfrak{C}_{\mathrm{ania}}$.

■

Appendix B. Generalization of Bach and Moulines (2013).

In this section, we give the demonstration of Theorem 8 which extends Theorem 1 from Bach and Moulines (2013); the demonstration is close to the original one.

B.1 Proof principle

For k in $\mathbb{N}^*$, the proof relies (1) on decomposing $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}\|^2]$ in two terms using the Minkowski inequality S6 to make appear a recursion $(\eta_k^0)_{k\in\mathbb{N}^*}$ without multiplicative noise, and another $(\alpha_k)_{k\in\mathbb{N}^*}$ without additive noise, (2) on an expansion of η_k^0 and $\overline{\eta}_k^0$ as polynomials in γ, and (3) on using the Hölder-type Assumption 2.2 to bound α_k. We define the sequence $(\eta_k^0)_{k\in\mathbb{N}^*}$ such that it involves only an additive noise:

$$\eta_k^0 = (\mathrm{I}_d - \gamma H_F)\eta_{k-1}^0 + \gamma \xi_k^{\mathrm{add}} . \tag{S11}$$

Then, we decompose $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}\|^2]$ in the following way using Minkowski inequality S6:

$$\mathbb{E}\left[\|H_F^{1/2}\overline{\eta}_{K-1}\|^2\right] \leq \left(\mathbb{E}\left[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2\right]^{1/2} + \mathbb{E}\left[\|H_F^{1/2}(\overline{\eta}_{K-1} - \overline{\eta}_{K-1}^0)\|^2\right]^{1/2}\right)^2 . \tag{S12}$$

The goal is then to establish a bound for the two above quantities.

1. Bounding $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2]$.

The bound on $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2]$ is given in Lemma S38. For k in $\mathbb{N}^*$, the proof relies on an expansion of η_k^0 and $\overline{\eta}_k^0$ as polynomials in γ. The recursion defining the sequence $(\eta_k^0)_{k\in\mathbb{N}^*}$ is $\eta_k^0 = (\mathrm{I}_d - \gamma H_F)\eta_{k-1}^0 + \gamma\xi_k^{\mathrm{add}}$. If we denote $M_i^k = (\mathrm{I}_d - \gamma H_F)^{k-i}$ and $M_i^{i-1} = \mathrm{I}_d$, we have:

$$\eta_k^0 = M_1^k \eta_0^0 + \gamma \sum_{i=1}^k M_{i+1}^k \xi_k^{\mathrm{add}} .$$

For K in $\mathbb{N}^*$, it leads to $\overline{\eta}_{K-1}^0 = \frac{1}{K}\sum_{k=0}^{K-1} M_1^k\eta_0^0 + \frac{\gamma}{K}\sum_{k=1}^{K-1}\left(\sum_{i=k}^K M_{k+1}^i\right)\xi_k^{\mathrm{add}}$, and with Minkowski inequality S6 to:

$$\mathbb{E}\left[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2\right]^{1/2} \leq \mathbb{E}\left[\left\|\frac{H_F^{1/2}}{K}\sum_{k=0}^{K-1} M_1^k\eta_0^0\right\|^2\right]^{1/2} + \mathbb{E}\left[\left\|\frac{\gamma H_F^{1/2}}{K}\sum_{k=1}^{K-1}\sum_{i=k}^{K} M_{k+1}^i \xi_k^{\mathrm{add}}\right\|^2\right]^{1/2} . \tag{S13}$$

The left term depends only on initial conditions η_0^0 $(= \eta_0)$ and the right term depends only on the additive noise. This is why, in the proof, we expend η_{k-1}^0 and $\overline{\eta}_{k-1}^0$ separately for the

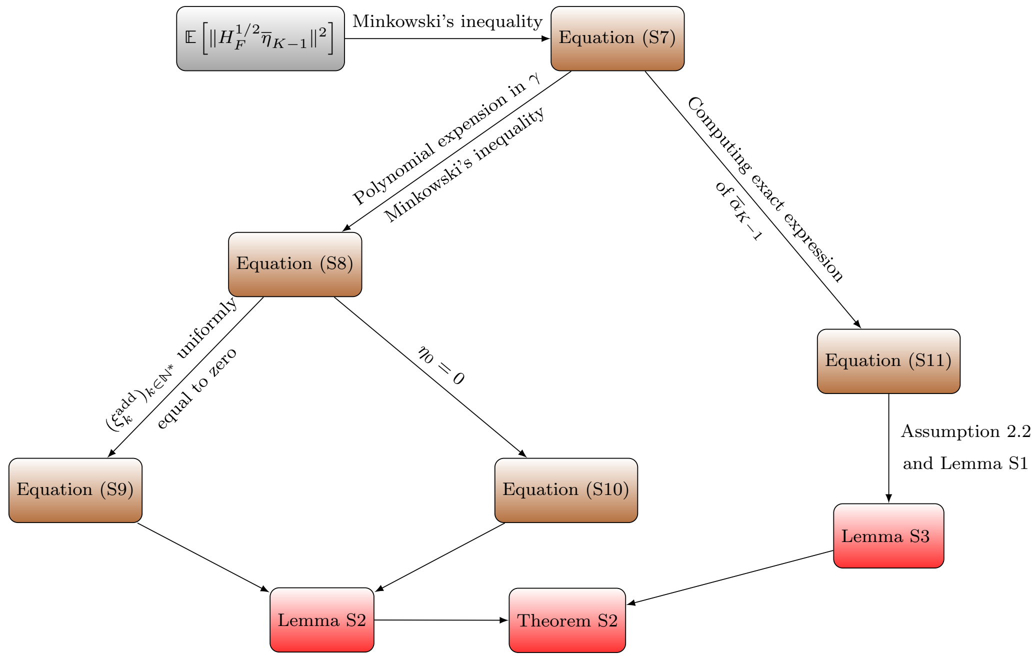


Figure S9: Proof principle of Theorem S41

noise process (i.e., when assuming $\eta_0 = 0$) and for the noise-free process that depends only on the initial conditions (i.e. when assuming that the additive noise $(\xi_k^{\mathrm{add}})_{k\in\mathbb{N}^*}$ is uniformly equal to zero). In the end, the two bounds computed separately may be added.

2. Bounding $\mathbb{E}[\|H_F^{1/2}(\overline{\eta}_{K-1} - \overline{\eta}_{K-1}^0)\|^2]$.

The bound on $\mathbb{E}[\|H_F^{1/2}(\overline{\eta}_{K-1} - \overline{\eta}_{K-1}^0)\|^2]$ is given in Lemma S39. For k in $\mathbb{N}^*$, the demonstration is based on an exact expression of $\alpha_k = \eta_k - \eta_k^0$ and $\overline{\alpha}_k$ computed by unrolling the recursion from α_k to α_0. Because $\alpha_0 = 0$ and because there is no additive noise involved in α_k, we obtain for K in $\mathbb{N}^*$, an expression of $\overline{\alpha}_{K-1}$ that depends only on the multiplicative noise at iteration k in $\{1, \cdots, K\}$:

$$\overline{\alpha}_{K-1} = \frac{\gamma}{K} \sum_{k=1}^{K-1} (\mathrm{I}_d - (\mathrm{I}_d - \gamma H_F)^{K-k})(\gamma H_F)^{-1} \xi_k^{\mathrm{mult}}(\eta_{k-1})\,.$$

We then show (Equation (S16)) that bounding $\mathbb{E}[\|H_F^{1/2}(\overline{\eta}_{K-1} - \overline{\eta}_{K-1}^0)\|^2]$ leads to bound the following sum $\frac{1}{K^2}\sum_{k=1}^{K-1} \mathbb{E}[\|H_F^{-1/2}\xi_k^{\mathrm{mult}}(\eta_{k-1})\|^2 \mid \mathcal{F}_{k-1}]$, and this bound is established using the Hölder-type Assumption 2.2; which concludes this part of the proof.

B.2 Two bounds

In this subsection, we give two lemmas that provide a bound on $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2]$ and $\mathbb{E}[\|H_F^{1/2}(\overline{\eta}_{K-1} - \overline{\eta}_{K-1}^0)\|^2]$.

These bounds are required due to the decomposition of $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}\|^2]$ done in Equation (S12).

- The bound on $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_k^0\|^2]$ is given in Lemma S38. It is established by decomposing the noise process and the noise-free process. The bound on the noise process comes from Lemma 2 (Bach and Moulines, 2013) and involves the additive noise's covariance $\mathfrak{C}_{\mathrm{ania}}$.
- The bound on $\mathbb{E}[\|H_F^{1/2}(\overline{\eta}_K - \overline{\eta}_K^0)\|^2]$ is established in Lemma S39.

Note that in order to demonstrate Lemma S39, we need to bound $\sum_{k=1}^K \|H_F^{1/2}\eta_k\|^2/K$. This is done in Lemma S35 which is an adaptation of Lemma 1 from Bach and Moulines (2013) to random mechanisms. This auxiliary lemma holds for any kind of multiplicative noise—linear or non-linear.
Below lemma provides a bound on $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_k^0\|^2]$.

Lemma S38 (Bound on $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_k^0\|^2]$) *Under the setting considered in Definition 2, under Assumption 1, for any K in $\mathbb{N}^*$ and any step-size $\gamma \in \mathbb{R}^+$ s.t. $\gamma R_F^2 \leq 1$, the sequence $(\eta_k^0)_{k\in\mathbb{N}^*}$ defined in Equation* (S11) *verifies the following bound:*

$$\mathbb{E}\left[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2\right]^{1/2} \leq \frac{1}{\sqrt{K}}\left(\frac{\|H_F^{-1/2}\eta_0\|}{\gamma\sqrt{K}} \wedge \frac{\|\eta_0\|}{\sqrt{\gamma}} + \sqrt{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)}\right).$$

Proof
The proof relies on the proof presented by Bach and Moulines (2013) and is done separately for the noise process and for the noise-free process that depends only on the initial condition. The bounds may then be added (see the discussion in Appendix B.1).
Noise-free process.
As in section A.3 from Bach and Moulines (2013), we assume in this section that the random fields $(\xi_k^{\mathrm{add}})_{k\in\mathbb{N}^*}$ is uniformly equal to zero and that $\gamma R_F^2 \leq 1$. We thus have for any k in $\mathbb{N}^*$ that $\eta_k^0 = (\mathrm{I}_d - \gamma H_F)\eta_{k-1}^0$.
First inequality. By recursion, we have $\eta_k^0 = (\mathrm{I}_d - \gamma H_F)^k\eta_0^0$, averaging over K in $\mathbb{N}^*$ and computing the resulting geometric sum, we have:

$$\overline{\eta}_{K-1}^0 = \frac{1}{K}\sum_{k=0}^{K-1}(\mathrm{I}_d - \gamma H_F)^k\eta_0^0 = \frac{1}{K}(\mathrm{I}_d - (\mathrm{I}_d - \gamma H_F)^{K-1})(\gamma H_F)^{-1}\eta_0^0 \preccurlyeq \frac{1}{\gamma K}H_F^{-1}\eta_0^0.$$

And because $\eta_0^0 = \eta_0$, it gives $\mathbb{E}\left[\langle\overline{\eta}_{K-1}^0, H_F\overline{\eta}_{K-1}^0\rangle\right] \leq \frac{\|H_F^{1/2}\eta_0\|^2}{\gamma^2K^2}$.
Second inequality. From the expression of η_k^0 flows:

$$\mathbb{E}[\|\eta_k^0\|^2] = \mathbb{E}[\|\eta_{k-1}^0\|^2] - 2\gamma\left\langle\eta_{k-1}^0, H_F\eta_{k-1}^0\right\rangle + \gamma^2\left\langle\eta_{k-1}^0, H_F^2\eta_{k-1}^0\right\rangle.$$

Considering that $H_F \preccurlyeq \mathrm{Tr}(H_F)\,\mathrm{I}_d \preccurlyeq R_F^2\mathrm{I}_d$ (Definition 2) and that $\gamma R_F^2 \leq 1$, because $\eta_0^0 = \eta_0$, by convexity we have: $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2] \leq \frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|H_F^{1/2}\eta_{k-1}^0\|^2] \leq \frac{\|\eta_0\|^2}{\gamma K}$.
Putting things together.
In the end, we take the minimum of the two above bounds and obtain that:

$$\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2] \leq \frac{\|H_F^{-1/2}\eta_0\|^2}{\gamma^2K^2} \wedge \frac{\|\eta_0\|^2}{\gamma K}. \tag{S14}$$

Noise process.

We assume in this part that $\eta_0^0 = \eta_0 = 0$. We apply Lemma 2 from Bach and Moulines (2013) to η_{k-1}^0. This sequence of iterates has an i.i.d. noise process $(\xi_k^{\mathrm{add}})_{k\in\mathbb{N}^*}$ which is such that $\mathbb{E}\left[\xi_k^{\mathrm{add}} \otimes \xi_k^{\mathrm{add}}\right] = \mathfrak{C}_{\mathrm{ania}}$ (existence guaranteed by Assumption 1). Therefore we have:

$$\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2] \leq \frac{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}} H_F^{-1}\right)}{K} . \tag{S15}$$

Putting things together. We now take results derived from the part without noise and the part with noise, and we get from Minkowski inequality:

$$\mathbb{E}\left[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2\right]^{1/2} \leq \frac{1}{\sqrt{K}}\left(\frac{\|H_F^{-1/2}\eta_0\|}{\gamma\sqrt{K}} \wedge \frac{\|\eta_0\|}{\sqrt{\gamma}} + \sqrt{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}} H_F^{-1}\right)}\right) .$$

■

Below lemma provides a bound on $\mathbb{E}[\|H_F^{1/2}(\overline{\eta}_K - \overline{\eta}_K^0)\|^2]$.

Lemma S39 (Bound on $\mathbb{E}[\|H_F^{1/2}(\overline{\eta}_K - \overline{\eta}_K^0)\|^2]$) *Under the setting considered in Definition 2 with $\mu > 0$, under Assumption 1 , under Assumptions 2.1 and 2.2, for any K in $\mathbb{N}^*$ and any step-size $\gamma \in \mathbb{R}^+$ s.t. $\gamma(R_F^2 + 2\mathcal{M}_2) < 1$, the sequence $(\overline{\eta}_k - \overline{\eta}_k^0)_{k\in\mathbb{N}^*}$ verifies the following bound:*

$$\begin{aligned}\mathbb{E}\left[\|H_F^{1/2}(\overline{\eta}_K - \overline{\eta}_K^0)\|^2\right]^{1/2}] \leq \frac{1}{\sqrt{K}}\Bigg(&\sqrt{\mathcal{M}_1\mu^{-1}}\left(\frac{5\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)}\right)^{1/4} \\ &+ \sqrt{\mathcal{M}_2\mu^{-1}}\left(\frac{15\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)}\right)^{1/2}\Bigg) .\end{aligned}$$

Remark S40 *To demonstrate Lemma S39, we use the Hölder-type Assumption 2.2. This is why we obtain a term with a square root in the bound.*

Proof

Let k in $\mathbb{N}^*$, we denote $\alpha_k = \eta_k - \eta_k^0$, with $\eta_k = (\mathrm{I}_d - \gamma H_F)\eta_{k-1} + \gamma\xi_k(\eta_{k-1})$ and $\eta_k^0 = (\mathrm{I}_d - \gamma H_F)\eta_{k-1}^0 + \gamma\xi_k^{\mathrm{add}}$. First, we write the exact expression of α_{k-1}:

$$\begin{aligned}\alpha_k &= (\mathrm{I}_d - \gamma H_F)\alpha_{k-1} + \gamma(\xi_k(\eta_{k-1}) - \xi_k^{\mathrm{add}}) \\ &= (\mathrm{I}_d - \gamma H_F)^k\alpha_0 + \gamma\sum_{i=1}^{k}(\mathrm{I}_d - \gamma H_F)^{k-i}(\xi_i(\eta_{i-1}) - \xi_i^{\mathrm{add}}) ,\end{aligned}$$

and because $\eta_0^0 = \eta_0$, it follows that $\alpha_0 = \eta_0 - \eta_0^0 = 0$. Averaging over K in $\mathbb{N}^*$, we have the exact expression of $\overline{\alpha}_{K-1}$:

$$\begin{aligned}\overline{\alpha}_{K-1} &= \frac{\gamma}{K}\sum_{k=0}^{K-1}\sum_{i=1}^{k}(\mathrm{I}_d - \gamma H_F)^{k-i}(\xi_i(\eta_{i-1}) - \xi_i^{\mathrm{add}})) \\ &= \frac{\gamma}{K}\sum_{i=1}^{K-1}\left(\sum_{k=i}^{K-1}(\mathrm{I}_d - \gamma H_F)^{k-i}\right)(\xi_i(\eta_{i-1}) - \xi_i^{\mathrm{add}})) .\end{aligned}$$

Computing the geometric sum results in:

$$\overline{\alpha}_{K-1} = \frac{\gamma}{K} \sum_{k=1}^{K-1} (\mathrm{I}_d - (\mathrm{I}_d - \gamma H_F)^{K-k})(\gamma H_F)^{-1}(\xi_k(\eta_{k-1}) - \xi_k^{\mathrm{add}})\,.$$

And because for any k in $\mathbb{N}$, $0 \preccurlyeq (\mathrm{I}_d - \gamma H_F)^k \preccurlyeq \mathrm{I}_d$, we obtain:

$$\overline{\alpha}_{K-1} \preccurlyeq \frac{1}{K} \sum_{k=1}^{K-1} H_F^{-1}(\xi_k(\eta_{k-1}) - \xi_k^{\mathrm{add}})\,,$$

hence $\|H_F^{1/2}\overline{\alpha}_{K-1}\|^2 = \|\frac{1}{K}\sum_{k=1}^{K-1} H_F^{-1/2}(\xi_k(\eta_{k-1}) - \xi_k^{\mathrm{add}})\|^2$. We take full expectation, because for any k in $\mathbb{N}^*$, by Definitions 2 and 4, $\xi_k^{\mathrm{mult}}(\eta_{k-1}) = \xi_k(\eta_{k-1}) - \xi_k^{\mathrm{add}}$ is $\mathcal{F}_k$-measurable and $\mathbb{E}\left[\xi_k^{\mathrm{mult}}(\eta_{k-1}) \;\middle|\; \mathcal{F}_{k-1}\right] = 0$, we can unroll the sum and we have in the end that the variance of the sum is the sum of variances:

$$\mathbb{E}\left[\left\|H_F^{1/2}\overline{\alpha}_{K-1}\right\|^2\right] \leq \frac{1}{K^2} \sum_{k=1}^{K-1} \mathbb{E}\left[\left\|H_F^{-1/2}\xi_k^{\mathrm{mult}}(\eta_{k-1})\right\|^2 \;\middle|\; \mathcal{F}_{k-1}\right]\,. \tag{S16}$$

Computing $\mathbb{E}[\|H_F^{-1/2}\xi_k^{\mathrm{mult}}(\eta_{k-1})\|^2 \mid \mathcal{F}_{k-1}]$ for k in $\mathbb{N}$, we first have:

$$\|H_F^{-1/2}\xi_k^{\mathrm{mult}}(\eta_{k-1})\|^2 \leq \|\!|H_F^{-1/2}|\!\|^2 \|\xi_k^{\mathrm{mult}}(\eta_{k-1})\|^2\,,$$

where we used Inequality 2. Because H_F is a symmetric semi-positive matrix, we have $\|\!|H_F^{-1/2}|\!\|^2 = 1/\mu$, hence: $\|H_F^{-1/2}\xi_k^{\mathrm{mult}}(\eta_{k-1})\|^2 \leq \mu^{-1}\|\xi_k^{\mathrm{mult}}(\eta_{k-1})\|^2$. Taking expectation conditionally to the σ-algebra $\mathcal{F}_{k-1}$ and invoking Assumption 2.2 gives:

$$\mathbb{E}[\|H_F^{-1/2}\xi_k^{\mathrm{mult}}(\eta_{k-1})\|^2 \mid \mathcal{F}_{k-1}] \leq \mu^{-1}(\mathcal{M}_1\|H_F^{1/2}\eta_{k-1}\| + 3\mathcal{M}_2\|H_F^{1/2}\eta_{k-1}\|^2)\,. \tag{S17}$$

Combining equations S16 and S17, we obtain:

$$\mathbb{E}[\|H_F^{1/2}\overline{\alpha}_{K-1}\|^2] \leq \frac{\mathcal{M}_1}{\mu K^2}\sum_{k=1}^{K-1}\mathbb{E}[\|H_F^{1/2}\eta_{k-1}\|] + \frac{3\mathcal{M}_2}{\mu K^2}\sum_{k=1}^{K-1}\mathbb{E}[\|H_F^{1/2}\eta_{k-1}\|^2]\,.$$

Now using Jensen's inequality for concave function allows us to write:

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|H_F^{1/2}(w - w_*)\|] \leq \frac{1}{K}\sum_{k=1}^{K}\sqrt{\mathbb{E}[\|H_F^{1/2}(w - w_*)\|^2]} \leq \sqrt{\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|H_F^{1/2}(w - w_*)\|^2]}\,,$$

thus we have:

$$\mathbb{E}[\|H_F^{1/2}\overline{\alpha}_{K-1}\|^2] \leq \frac{\mathcal{M}_1}{\mu K}\sqrt{\frac{1}{K}\sum_{k=1}^{K-1}\mathbb{E}[\|H_F^{1/2}\eta_{k-1}\|^2]} + \frac{3\mathcal{M}_2}{\mu K^2}\sum_{k=1}^{K-1}\mathbb{E}[\|H_F^{1/2}\eta_{k-1}\|^2]\,.$$

Using Lemma S35 (with $\eta_0 = 0$), we finally obtain:

$$\mathbb{E}[\|H_F^{1/2}\overline{\alpha}_{K-1}\|^2] \leq \frac{1}{K}\left(\mathcal{M}_1\mu^{-1}\sqrt{\frac{5\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)}} + \frac{15\mathcal{A}\gamma\mathcal{M}_2\mu^{-1}}{1-\gamma(R_F^2+2\mathcal{M}_2)}\right).$$

In the end, we take the square root (and use that for any a, b in $\mathbb{R}_+$, $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$) which allows concluding:

$$\mathbb{E}\left[\|H_F^{1/2}(\overline{\eta}_K - \overline{\eta}_K^0)\|^2\right]^{1/2} \leq \frac{1}{\sqrt{K}}\left(\sqrt{\mathcal{M}_1\mu^{-1}}\left(\frac{5\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)}\right)^{1/4} + \sqrt{\mathcal{M}_2\mu^{-1}}\left(\frac{15\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)}\right)^{1/2}\right).$$

■

B.3 Final theorem

In this section, we gather the pieces of proof required to demonstrate Theorem 8.

Theorem S41 (Non-linear multiplicative noise) *Under Assumptions 1 and 2, considering any constant step-size γ such that $\gamma(R_F^2+2\mathcal{M}_2) \leq 1/2$, then for any K in $\mathbb{N}^*$, the sequence $(w_k)_{k\in\mathbb{N}^*}$ produced by a setting such as in Definition 2 verifies the following bound:*

$$\mathbb{E}[F(\overline{w}_{K-1}) - F(w_*)] \leq \frac{1}{2K}\left(\frac{\|H_F^{-1/2}\eta_0\|}{\gamma\sqrt{K}} \wedge \frac{\|\eta_0\|}{\sqrt{\gamma}} + \sqrt{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)} + (10\mathcal{A}\gamma)^{1/4}\sqrt{\mathcal{M}_1\mu^{-1}} + (30\mathcal{A}\gamma)^{1/2}\sqrt{\mathcal{M}_2\mu^{-1}}\right)^2.$$

Proof

As explained in the discussion in Appendix B.1 (Equation (S12)), we define the sequence $(\eta_k^0)_{k\in\mathbb{N}^*}$ which involves only an additive noise $\eta_k^0 = (\mathrm{I}_d - \gamma H_F)\eta_{k-1}^0 + \gamma\xi_k^{\mathrm{add}}$. Then, we decompose $\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}\|]$ using Minkowski's inequality S6:

$$\mathbb{E}\left[\|H_F^{1/2}\overline{\eta}_{K-1}\|^2\right] \leq \left(\mathbb{E}\left[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2\right]^{1/2} + \mathbb{E}\left[\|H_F^{1/2}(\overline{\eta}_{K-1} - \overline{\eta}_{K-1}^0)\|^2\right]^{1/2}\right)^2. \quad \text{(S18)}$$

First term.

To bound the first term, we use Lemma S38 which gives:

$$\mathbb{E}\left[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2\right]^{1/2} \leq \frac{1}{\sqrt{K}}\left(\frac{\|H_F^{-1/2}\eta_0\|}{\gamma\sqrt{K}} \wedge \frac{\|\eta_0\|}{\sqrt{\gamma}} + \sqrt{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)}\right).$$

Second term.

From Lemma S39, we have:

$$\mathbb{E}\left[\|H_F^{1/2}(\overline{\eta}_K - \overline{\eta}_K^0)\|^2\right]^{1/2} \leq \frac{1}{\sqrt{K}}\left(\sqrt{\mathcal{M}_1\mu^{-1}}\left(\frac{5\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)}\right)^{1/4} + \sqrt{\mathcal{M}_2\mu^{-1}}\left(\frac{15\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)}\right)^{1/2}\right).$$

Final bound. Hence, back to Equation (S18), we get:

$$\mathbb{E}\left[\|H_F^{1/2}\overline{\eta}_{K-1}\|^2\right]^{1/2} \leq \frac{1}{\sqrt{K}}\left(\frac{\|H_F^{-1/2}\eta_0\|}{\gamma\sqrt{K}} \wedge \frac{\|\eta_0\|}{\sqrt{\gamma}} + \sqrt{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)} + \sqrt{\mathcal{M}_1\mu^{-1}}\left(\frac{5\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)}\right)^{1/4} + \sqrt{\mathcal{M}_2\mu^{-1}}\left(\frac{15\mathcal{A}\gamma}{1-\gamma(R_F^2+2\mathcal{M}_2)}\right)^{1/2}\right),$$

and considering $\gamma(R_F^2+2\mathcal{M}_2) \leq 1/2$, it concludes the proof because $\mathbb{E}[F(\overline{w}_{K-1}) - F(w_*)] = \mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}\|^2]/2$:

$$\mathbb{E}[F(\overline{w}_{K-1}) - F(w_*)] \leq \frac{1}{2K}\left(\frac{\|H_F^{-1/2}\eta_0\|}{\gamma\sqrt{K}} \wedge \frac{\|\eta_0\|}{\sqrt{\gamma}} + \sqrt{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}}H_F^{-1}\right)} + (10\mathcal{A}\gamma)^{1/4}\sqrt{\mathcal{M}_1\mu^{-1}} + (30\mathcal{A}\gamma)^{1/2}\sqrt{\mathcal{M}_2\mu^{-1}}\right)^2.$$

■

Appendix C. Generalisation of Bach and Moulines (2013) for linear multiplicative noise.

In this Section, we give the demonstration of Theorem 12 which extends Theorem 1 from Bach and Moulines (2013) to the case of linear multiplicative noise. The demonstration follows the same steps as the one given by Bach and Moulines (2013). The minor differences lie in the generality of the form of the multiplicative noise in our approach. Bach and Moulines (2013) only analyse LMS algorithm, while we here consider (LSA) with assumptions on the linear multiplicative noise process. Moreover, our theorem decomposes into 3 terms instead of 2.

C.1 Proof principle

For k in $\mathbb{N}^*$, the proof relies on an expansion of η_k and $\overline{\eta}_k$ as polynomials in γ. Because we consider a linear multiplicative noise, there exists a matrix Ξ_k in $\mathbb{R}^{d\times d}$ s.t. for any z in $\mathbb{R}^d$,

$\xi_k^{\text{mult}}(z) = \Xi_k z$ (Assumption 3); hence the recursion defined in Definition 2 can be rewritten as:

$$\eta_k = \eta_{k-1} - \gamma \nabla F(\eta_{k-1}) + \gamma \xi_k^{\text{mult}}(\eta_{k-1}) + \gamma \xi_k^{\text{add}} = (\mathrm{I}_d - \gamma H_F + \gamma \Xi_k)\eta_{k-1} + \gamma \xi_k^{\text{add}} .$$

We denote $M_i^k = (\mathrm{I}_d - \gamma H_F + \gamma \Xi_k) \cdots (\mathrm{I}_d - \gamma H_F + \gamma \Xi_i)$ and $M_i^{i-1} = \mathrm{I}_d$, then we have that $\eta_k = M_1^k \eta_0 + \gamma \sum_{i=1}^k M_{i+1}^k \xi_k^{\text{add}}$.
For K in $\mathbb{N}^*$, it leads to $\overline{\eta}_{K-1} = \frac{1}{K} \sum_{k=0}^{K-1} M_1^k \eta_0 + \frac{\gamma}{K} \sum_{k=1}^{K-1} \left(\sum_{i=k}^{K} M_{k+1}^i \right) \xi_k^{\text{add}}$, and with Minkowski's inequality S6 to:

$$\sqrt{\mathbb{E}\left[\left\| H_F^{1/2} \overline{\eta}_{K-1} \right\|^2 \right]} \leq \mathbb{E}\left[\left\| \frac{H_F^{1/2}}{K} \sum_{k=0}^{K-1} M_1^k \eta_0 \right\|^2 \right]^{1/2} + \mathbb{E}\left[\left\| \frac{\gamma H_F^{1/2}}{K} \sum_{k=1}^{K-1} \sum_{i=k}^{K} M_{k+1}^i \xi_k^{\text{add}} \right\|^2 \right]^{1/2} . \tag{S19}$$

The left term depends only on initial conditions and the right term depends only on the noise process.
This is why, in the proof, we expend η_{k-1} and $\overline{\eta}_{k-1}$ separately for the noise process (i.e., when assuming $\eta_0 = 0$) and for the noise-free process that depends only on the initial conditions (i.e. when assuming that the additive noise $(\xi_k^{\text{add}})_{k \in \mathbb{N}^*}$ is uniformly equal to zero). In the end, the two bounds computed separately may be added.
To study the noise process, inspiring from Bach and Moulines (2013), we define the following sequence:

$$\begin{cases} \eta_k^0 = (\mathrm{I}_d - \gamma H_F)\eta_{k-1}^0 + \gamma \xi_k^{\text{add}} \\ \eta_k^r = (\mathrm{I}_d - \gamma H_F)\eta_{k-1}^r + \gamma \xi_k^{\text{mult}}\left(\eta_{k-1}^{r-1}\right) \qquad \text{with} \qquad \forall r \geq 0 \,, \eta_0^r = 0 \,. \end{cases} \tag{S20}$$

Then, we decompose $\mathbb{E}[\| H_F^{1/2} \overline{\eta}_{K-1} \|^2]$ in the following way using Minkowski's inequality S6:

$$\sqrt{\mathbb{E}\left[\| H_F^{1/2} \overline{\eta}_{K-1} \|^2 \right]} \leq \mathbb{E}[\| H_F^{1/2} \sum_{i=0}^{r} \overline{\eta}_{K-1}^i \|^2]^{1/2} + \mathbb{E}[\| H_F^{1/2} (\overline{\eta}_{K-1} - \sum_{i=0}^{r} \overline{\eta}_{K-1}^i) \|^2]^{1/2}.$$

The goal is then to establish a bound for the two above quantities.

C.2 Lemmas for the noise process

In this Subsection, we provide lemmas for the noise process, and thus we suppose that $\eta_0 = 0$. The noise-free process is later considered in Appendix C.3 and puts together with the results of the coming Subsection. The sketch of the proof relies on establishing two bounds.

- For r, k in $\mathbb{N} \times \mathbb{N}^*$, noting $\alpha_k^r = \eta_k - \sum_{i=0}^r \eta_k^i$, the first one is a bound on $\mathbb{E}[\| H_F^{1/2} \overline{\alpha}_{K-1}^r \|^2]$ that tends to zero when r tends to $+\infty$.
- The second one is on $\sum_{i=0}^r \mathbb{E}[\| H_F^{1/2} \overline{\eta}_{K-1}^i \|^2]$ and is established using Lemma 2 from (Bach and Moulines, 2013). It will correspond to the final variance term and it involves the additive noise's covariance $\mathfrak{C}_{\text{ania}}$.

In the following, we provide Lemmas S42 to S44. Let r, k in $\mathbb{N} \times \mathbb{N}^*$.

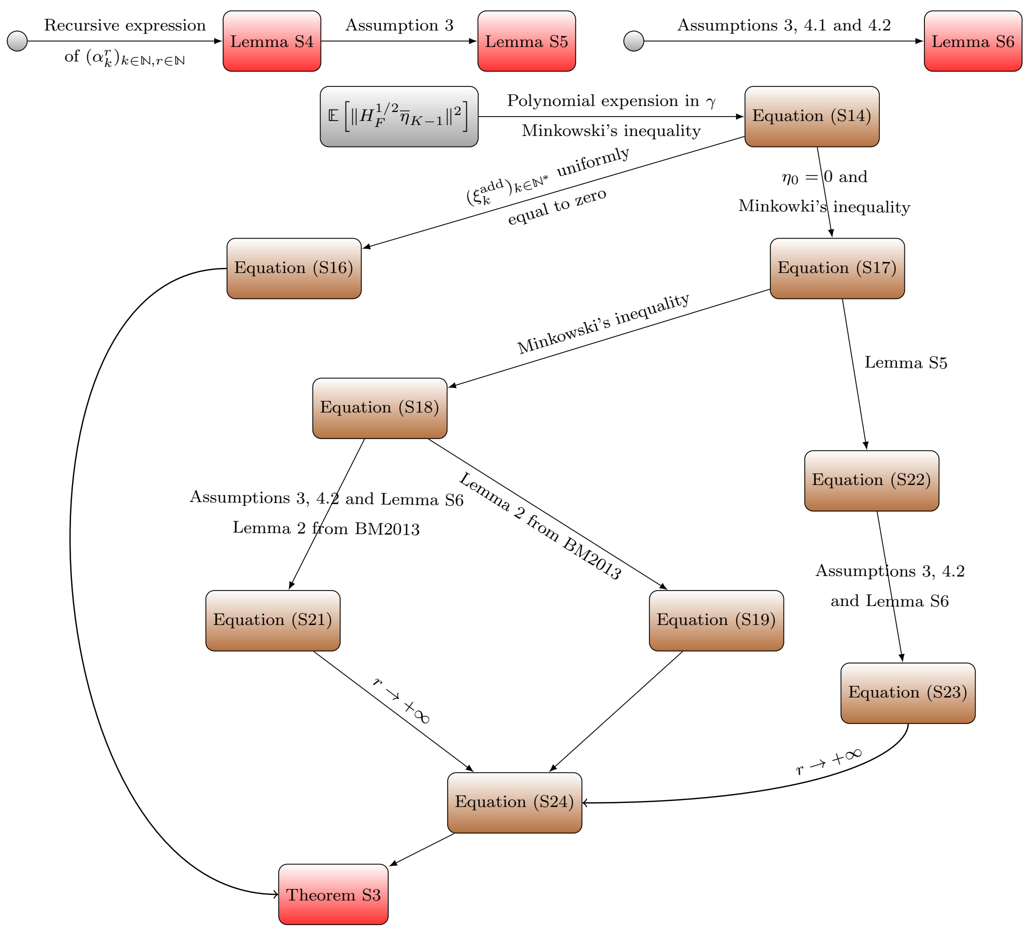


Figure S10: Proof principle of Theorem S45.

- Lemma S42 builds a recursive expression of $\alpha_k^r = \eta_k - \sum_{i=0}^r \eta_k^i$.
- Lemma S43 provides a bound on $\mathbb{E}[\|H_F^{1/2}\overline{\alpha}_{K-1}^r\|^2]$ which involves $\mathbb{E}\|\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^r\right)\|^2$.
- Lemma S44 bounds the covariance of η_{k-1}^r, this result will be necessary when computing the expectation of $\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^r\right)^{\otimes 2}$.

Below, we provide the lemma that builds a recursive expression of $\eta_k - \sum_{i=0}^r \eta_k^i$, with k, r in $\mathbb{N}^*$.

Lemma S42 (A recursion on $\eta_k - \sum_{i=0}^r \eta_k^i$) *Under the setting given in Definition 2, considering that $\xi_k^{\mathrm{mult}}(\cdot)$ is linear (Assumption 3), for any k in $\mathbb{N}^*$ and any step-size $\gamma > 0$, considering $(\eta_k^r)_{r\in\mathbb{N}}$ as given by Equation* (S20), *denoting for r in $\mathbb{N}$, $\alpha_k^r = \eta_k - \sum_{i=0}^r \eta_k^i$, we have the following recursive expression for the sequence of iterate $(\alpha_k^r)_{r\in\mathbb{N}}$:*

$$\forall r \geq 0, \alpha_k^r = (\mathrm{I}_d - \gamma H_F)\alpha_{k-1}^r + \xi_k^{\mathrm{mult}}\left(\alpha_{k-1}^r\right) + \gamma \xi_k^{\mathrm{mult}}\left(\eta_{k-1}^r\right) .$$

Proof Let k in $\mathbb{N}^*$, the proof is done by recursion. For $r = 0$, by Definitions 2 and 4, we have $\eta_k = \eta_{k-1} - \gamma \nabla F(w_{k-1}) + \gamma \xi_k(\eta_{k-1}) = (\mathrm{I}_d - \gamma H_F)\eta_{k-1} + \gamma \xi_k^{\mathrm{add}} + \gamma \xi_k^{\mathrm{mult}}(\eta_{k-1})$, which gives:

$$\begin{aligned}
\alpha_k^0 = \eta_k - \eta_k^0 &= \left\{(\mathrm{I}_d - \gamma H_F)\eta_{k-1} + \gamma \xi_k^{\mathrm{add}} + \gamma \xi_k^{\mathrm{mult}}(\eta_{k-1})\right\} - \left\{(\mathrm{I}_d - \gamma H_F)\eta_{k-1}^0 + \gamma \xi_k^{\mathrm{add}}\right\} \\
&= (\mathrm{I}_d - \gamma H_F)(\eta_{k-1} - \eta_{k-1}^0) + \gamma \xi_k^{\mathrm{mult}}(\eta_{k-1}) \\
&= (\mathrm{I}_d - \gamma H_F)(\eta_{k-1} - \eta_{k-1}^0) + \gamma \xi_k^{\mathrm{mult}}\left(\eta_{k-1} - \eta_{k-1}^0\right) + \gamma \xi_k^{\mathrm{mult}}\left(\eta_{k-1}^0\right) ,
\end{aligned}$$

which is possible because ξ_k^{mult} is linear (Assumption 3). To go from r to $r+1$, we have $\alpha_k^{r+1} = \eta_k - \sum_{i=0}^{r+1} \eta_k^i = \eta_k - \sum_{i=0}^r \eta_k^i - \eta_k^{r+1}$. Then by definition of η_k^{r+1} and using the hypothesis:

$$\begin{aligned}
\alpha_k^{r+1} &= (\mathrm{I}_d - \gamma H_F)\left(\eta_{k-1} - \sum_{i=0}^r \eta_{k-1}^i\right) + \xi_k^{\mathrm{mult}}\left(\eta_{k-1} - \sum_{i=0}^r \eta_{k-1}^i\right) + \gamma \xi_k^{\mathrm{mult}}\left(\eta_{k-1}^r\right) \\
&\quad - (\mathrm{I}_d - \gamma H_F)\eta_{k-1}^{r+1} - \gamma \xi_k^{\mathrm{mult}}\left(\eta_{k-1}^r\right) \\
&= (\mathrm{I}_d - \gamma H_F)\left(\eta_{k-1} - \sum_{i=0}^{r+1} \eta_{k-1}^i\right) + \xi_k^{\mathrm{mult}}\left(\eta_{k-1} - \sum_{i=0}^{r+1} \eta_{k-1}^i\right) + \gamma \xi_k^{\mathrm{mult}}\left(\eta_{k-1}^{r+1}\right) ,
\end{aligned}$$

again by linearity. This concludes the proof. ■

The next lemma is the adaptation to our settings of Lemma 1 from Bach and Moulines (2013). We give a bound on $\mathbb{E}[\|H_F^{1/2}\overline{\alpha}_{K-1}^r\|^2]$ with a quantity that tends to 0. This result will be used in the final demonstration of Theorem S45.

Lemma S43 (Bound on $\eta_K - \sum_{i=0}^r \eta_K^i$) *Under the setting given in Definition 2, considering that ξ_k^{mult} is linear (Assumption 3), for any r, K in $\mathbb{N} \times \mathbb{N}^*$ and any step-size γ s.t. $\gamma(R_F^2 + \mathcal{M}_2) \leq 1$, the recursion $\alpha_K^r = \eta_K - \sum_{i=0}^r \eta_K^i$ verifies the following bound:*

$$\forall r \geq 0, (1 - \gamma(R_F^2 + \mathcal{M}_2))\mathbb{E}\left\langle \overline{\alpha}_{K-1}^r, H_F \overline{\alpha}_{K-1}^r \right\rangle \leq \frac{\gamma}{K} \sum_{k=1}^K \mathbb{E}\|\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^r\right)\|^2 .$$

Proof Let r, k in $\mathbb{N} \times \mathbb{N}^*$, we denote $\alpha_k^r = \eta_k - \sum_{i=0}^r \eta_k^i$, then we have shown in Lemma S42 that:

$$\alpha_k^r = (\mathrm{I}_d - \gamma H_F)\alpha_{k-1}^r + \xi_k^{\mathrm{mult}}\left(\alpha_{k-1}^r\right) + \gamma \xi_k^{\mathrm{mult}}\left(\eta_{k-1}^r\right) .$$

Taking the squared norm and developing it:

$$\begin{aligned}\|\alpha_k^r\|^2 = \left\|\alpha_{k-1}^r\right\|^2 + 2\gamma \left\langle \alpha_{k-1}^r, \xi_k^{\mathrm{mult}}\left(\alpha_{k-1}^r\right) + \xi_k^{\mathrm{mult}}\left(\eta_{k-1}^r\right) - H_F \alpha_{k-1}^r \right\rangle \\ + \gamma^2 \|\xi_k^{\mathrm{mult}}\left(\alpha_{k-1}^r\right) + \xi_k^{\mathrm{mult}}\left(\eta_{k-1}^r\right) - H_F \alpha_{k-1}^r\|^2 ,\end{aligned}$$

and developing the last term with Inequality 1 leads to:

$$\begin{aligned}\|\alpha_k^r\|^2 \leq \left\|\alpha_{k-1}^r\right\|^2 + 2\gamma \left\langle \alpha_{k-1}^r, \xi_k^{\mathrm{mult}}\left(\alpha_{k-1}^r\right) + \xi_k^{\mathrm{mult}}\left(\eta_{k-1}^r\right) - H_F \alpha_{k-1}^r \right\rangle \\ + 2\gamma^2 \left\{ \|\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^r\right)\|^2 + \|H_F \alpha_{k-1}^r - \xi_k^{\mathrm{mult}}\left(\alpha_{k-1}^r\right)\|^2 \right\} .\end{aligned}$$

Because α_{k-1}^r is $\mathcal{F}_{k-1}$-measurable and $\mathbb{E}[\xi_k^{\mathrm{mult}}\left(\alpha_{k-1}^r\right) \mid \mathcal{F}_{k-1}] = 0$ (expectation of $\xi_k^{\mathrm{mult}}(\cdot)$ is zero, see Definitions 2 and 4), taking expectation w.r.t. the σ-algebra $\mathcal{F}_{k-1}$, using Assumption 3 and again Definition 2 gives:

$$\begin{aligned}\mathbb{E}[\|H_F \alpha_{k-1}^r - \xi_k^{\mathrm{mult}}\left(\alpha_{k-1}^r\right)\|^2 \mid \mathcal{F}_{k-1}] &= \mathbb{E}[\|H_F \alpha_{k-1}^r\|^2 \mid \mathcal{F}_{k-1}] \\ &\quad + \mathbb{E}[\|\xi_k^{\mathrm{mult}}\left(\alpha_{k-1}^r\right)\|^2 \mid \mathcal{F}_{k-1}] \\ &\leq (R_F^2 + \mathcal{M}_2)\|H_F^{1/2}\alpha_{k-1}^r\|^2 .\end{aligned}$$

Hence:

$$\begin{aligned}\mathbb{E}[\|\alpha_k^r\|^2 \mid \mathcal{F}_{k-1}] \leq \|\alpha_{k-1}^r\|^2 - 2\gamma(1 - \gamma(R_F^2 + \mathcal{M}_2)) \left\langle \alpha_{k-1}^r, H_F \alpha_{k-1}^r \right\rangle \\ + 2\gamma^2 \mathbb{E}[\|\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^r\right)\|^2 \mid \mathcal{F}_{k-1}] ,\end{aligned}$$

which gives when taking full expectation and averaging over K in $\mathbb{N}^*$:

$$\begin{aligned}(1 - \gamma(R_F^2 + \mathcal{M}_2)) \frac{1}{K} \sum_{k=1}^K \mathbb{E}\left\langle \alpha_{k-1}^r, H_F \alpha_{k-1}^r \right\rangle \leq \frac{1}{2\gamma}(\|\alpha_0^r\|^2 - \left\|\alpha_{k-1}^r\right\|^2) \\ + \frac{\gamma}{K} \sum_{k=1}^K \mathbb{E}[\|\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^r\right)\|^2] ,\end{aligned}$$

and by convexity $\left\langle \overline{\alpha}_{K-1}^r, H \overline{\alpha}_{K-1}^r \right\rangle \leqslant \frac{1}{K} \sum_{k=1}^K \left\langle \alpha_{k-1}^r, H_F \alpha_{k-1}^r \right\rangle$, which allows to conclude as $\alpha_0^r = 0$. ■

In below lemma, we bound $\mathbb{E}\left[\eta_{k-1}^r \otimes \eta_{k-1}^r\right]$ for r, k in $\mathbb{N} \times \mathbb{N}^*$. It is required because we will use Lemma 2 from Bach and Moulines (2013) and apply it to the sequence $(\eta_{k-1}^r)_{k \in \mathbb{N}^*, r \in \mathbb{N}}$. The noise process of this sequence is equal to $\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^{r-1}\right)$; and computing the expectation of its covariance involves knowing $\mathbb{E}\left[\eta_{k-1}^r \otimes \eta_{k-1}^r\right]$.

Lemma S44 (Bounding the covariance of η_{k-1}^r) *Under the setting in Definition 2, under Assumptions 1, 3 and 4, i.e. considering that $\xi_k^{\mathrm{mult}}(\cdot)$ is linear, for any K in $\mathbb{N}^*$, any step-size $\gamma > 0$, and for any $r \geq 0$, we have the following bound on the covariance of η_{k-1}^r:*

$$\mathbb{E}\left[\eta_{k-1}^r \otimes \eta_{k-1}^r\right] \preccurlyeq \gamma^{r+1} Ш_{\mathrm{add}} Ш_{\mathrm{mult}}^r \mathrm{I}_d \,.$$

Proof

Let $r > 0$, we first prove by recursion that we have:

$$\forall k > 0 \,, \eta_k^{r+1} = \gamma \sum_{i=1}^{k} (\mathrm{I}_d - \gamma H_F)^{k-i} \xi_i^{\mathrm{mult}}(\eta_{i-1}^r) \,.$$

For $k = 0$, we indeed have $\eta_0^{r+1} = 0$. To go from k to $k+1$:

$$\begin{aligned}
\eta_{k+1}^{r+1} &= (\mathrm{I}_d - \gamma H_F)\eta_k^{r+1} + \gamma \xi_{k+1}^{\mathrm{mult}}(\eta_k^r) \qquad \text{by definition,} \\
&= \gamma \sum_{i=1}^{k} (\mathrm{I}_d - \gamma H_F)^{k-i} \xi_i^{\mathrm{mult}}(\eta_{i-1}^r) + \gamma (\mathrm{I}_d - \gamma H_F)^{(k+1)-(k+1)} \xi_{k+1}^{\mathrm{mult}}(\eta_k^r) \,,
\end{aligned}$$

by hypothesis, which allows concluding.

We now prove by recursion the main result of the lemma.

Initialization. For $r = 0$, by definition, we have $\eta_k^0 = (\mathrm{I}_d - \gamma H_F)\eta_{k-1}^0 + \gamma \xi_k^{\mathrm{add}}$, unrolling the sum gives $\eta_k^0 = (\mathrm{I}_d - \gamma H_F)^k \eta_0^0 + \gamma \sum_{i=1}^k (\mathrm{I}_d - \gamma H_F)^{k-i} \xi_i^{\mathrm{add}}$. Because we consider $\eta_0^0 = 0$ and given that the sequence of noise $(\xi_i^{\mathrm{add}})_{i \in [\![1,k]\!]}$ is independent at each iterations, we have:

$$\mathbb{E}\left[\eta_k^0 \otimes \eta_k^0\right] = \gamma^2 \sum_{i=1}^{k} (\mathrm{I}_d - \gamma H_F)^{k-i} \mathbb{E}\left[\xi_i^{\mathrm{add}} \otimes \xi_i^{\mathrm{add}}\right] (\mathrm{I}_d - \gamma H_F)^{k-i} \,.$$

Because the sequence of additive noise $(\xi_i^{\mathrm{add}})_{i \in \mathbb{N}^*}$is i.i.d., for any i in $\{1, \cdots, k\}$, we have that $\mathbb{E}\left[\xi_i^{\mathrm{add}} \otimes \xi_i^{\mathrm{add}}\right] = \mathfrak{C}_{\mathrm{ania}} \preccurlyeq Ш_{\mathrm{add}} H_F$ (Assumption 4.1), hence:

$$\mathbb{E}\left[\eta_k^0 \otimes \eta_k^0\right] \preccurlyeq \gamma^2 \sum_{i=1}^{k} (\mathrm{I}_d - \gamma H_F)^{k-i} Ш_{\mathrm{add}} H_F (\mathrm{I}_d - \gamma H_F)^{k-i} \,.$$

These matrices commute:

$$\begin{aligned}
\mathbb{E}\left[\eta_k^0 \otimes \eta_k^0\right] &\preccurlyeq \gamma^2 Ш_{\mathrm{add}} \sum_{i=1}^{k} (\mathrm{I}_d - \gamma H_F)^{2k-2i} H_F \,, \text{ and because it is a geometric sum:} \\
&\preccurlyeq \gamma^2 Ш_{\mathrm{add}} \left(\mathrm{I}_d - (\mathrm{I}_d - \gamma H_F)^{2k-2}\right) \left(\mathrm{I}_d - (\mathrm{I}_d - \gamma H_F)^2\right)^{-1} H_F \\
&\preccurlyeq \gamma^2 Ш_{\mathrm{add}} \left(\mathrm{I}_d - (\mathrm{I}_d - \gamma H_F)^{2k-2}\right) \left(2\gamma H_F - \gamma^2 H_F^2\right)^{-1} H_F \\
&\preccurlyeq \gamma Ш_{\mathrm{add}} H_F^{-1} H_F \qquad \text{because } \gamma H_F \preccurlyeq \mathrm{I}_d, \\
&\preccurlyeq \gamma Ш_{\mathrm{add}} \mathrm{I}_d \,.
\end{aligned}$$

Recursion. Let $r \geq 0$, to go from r to $r+1$, we start writing:

$$\eta_k^{r+1} \otimes \eta_k^{r+1} = \gamma^2 \sum_{i=1}^{k} (\mathrm{I}_d - \gamma H_F)^{k-1-i} \xi_i^{\mathrm{mult}}(\eta_{i-1}^r) \otimes \xi_i^{\mathrm{mult}}(\eta_{i-1}^r) (\mathrm{I}_d - \gamma H_F)^{k-1-i} \,.$$

Now we use linearity of the multiplicative noise (Assumption 3), thus there exists a matrix Ξ_k in $\mathbb{R}^{d\times d}$ s.t. for any z in $\mathbb{R}^d$, we have $\xi_k^{\mathrm{mult}}(z) = \Xi_k z$, and it leads to:

$$\eta_k^{r+1} \otimes \eta_k^{r+1} = \gamma^2 \sum_{i=1}^{k} (\mathrm{I}_d - \gamma H_F)^{k-i} \Xi_i (\eta_{i-1}^r \otimes \eta_{i-1}^r) \Xi_i^\top (\mathrm{I}_d - \gamma H_F)^{k-i} \,.$$

Taking full expectation, we have:

$$\begin{aligned}
\mathbb{E}\left[\eta_k^{r+1} \otimes \eta_k^{r+1}\right] &= \gamma^2 \sum_{i=1}^{k} (\mathrm{I}_d - \gamma H_F)^{k-i} \mathbb{E}\left[\mathbb{E}\left[\Xi_i (\eta_{i-1}^r \otimes \eta_{i-1}^r) \Xi_i^\top \;\middle|\; \sigma(\Xi_i)\right]\right] (\mathrm{I}_d - \gamma H_F)^{k-i} \\
&= \gamma^2 \sum_{i=1}^{k} (\mathrm{I}_d - \gamma H_F)^{k-i} \mathbb{E}\left[\Xi_i \mathbb{E}[\eta_{i-1}^r \otimes \eta_{i-1}^r \mid \sigma(\Xi_i)] \Xi_i^\top\right] (\mathrm{I}_d - \gamma H_F)^{k-i} \,,
\end{aligned}$$

and because for any i in $\{1, \cdots, k\}$, η_{i-1}^r is independent of Ξ_i, we have $\mathbb{E}\left[\eta_{i-1}^r \otimes \eta_{i-1}^r \;\middle|\; \sigma(\Xi_i)\right] = \mathbb{E}\left[\eta_{i-1}^r \otimes \eta_{i-1}^r\right] \preccurlyeq \gamma^{r+1} Ш_{\mathrm{add}} Ш_{\mathrm{mult}}^r \mathrm{I}_d$, where we use the hypothesis for r. We have in the end:

$$\mathbb{E}\left[\eta_k^{r+1} \otimes \eta_k^{r+1}\right] \preccurlyeq \gamma^{r+3} Ш_{\mathrm{add}} Ш_{\mathrm{mult}}^r \sum_{i=1}^{k} (\mathrm{I}_d - \gamma H_F)^{k-i} \mathbb{E}\left[\Xi_i \Xi_i^\top\right] (\mathrm{I}_d - \gamma H_F)^{k-i} \,.$$

Furthermore, by Assumption 4.2 we have $\mathbb{E}\left[\Xi_i \Xi_i^\top\right] \preccurlyeq Ш_{\mathrm{mult}} H_F$, thus:

$$\begin{aligned}
\mathbb{E}\left[\eta_k^{r+1} \otimes \eta_k^{r+1}\right] &\preccurlyeq \gamma^{r+3} Ш_{\mathrm{add}} Ш_{\mathrm{mult}}^{r+1} \sum_{i=1}^{k} (\mathrm{I}_d - \gamma H_F)^{2k-2-2i} H_F \\
&\preccurlyeq \gamma^{r+3} Ш_{\mathrm{add}} Ш_{\mathrm{mult}}^{r+1} \gamma^{-1} H_F^{-1} H_F \,,
\end{aligned}$$

because $\sum_{i=1}^{k} (\mathrm{I}_d - \gamma H_F)^{2k-2-2i} = \left(\mathrm{I}_d - (\mathrm{I}_d - \gamma H_F)^{2k}\right) \left(2\gamma H_F - \gamma^2 H_F^2\right)^{-1} \preccurlyeq \gamma^{-1} H_F^{-1}$. In the end, we have $\mathbb{E}[\eta_k^{r+1} \otimes \eta_k^{r+1}] \preccurlyeq \gamma^{r+2} Ш_{\mathrm{add}} Ш_{\mathrm{mult}}^{r+1} \mathrm{I}_d$, which concludes the proof.

■

C.3 Final theorem

In this section, we gather the pieces of proof required to demonstrate Theorem 12. As done in Appendix B, we consider separately the noise process and the noise-free process, then put them together to obtain the final result.

Theorem S45 (Linear multiplicative noise, convex case) *Under Assumption 1, under Assumptions 3 and 4 i.e. with a linear multiplicative noise, considering any constant step-size γ such that $\gamma(R_F^2 + \mathcal{M}_2) \leq 1$ and $4\gamma Ш_{\mathrm{mult}} R_F^2 \leq 1$, then for any K in $\mathbb{N}^*$, the sequence $(w_k)_{k \in \mathbb{N}^*}$ produced by a setting such as in Definition 2, verifies the following bound:*

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

Proof Let K in $\mathbb{N}^*$, the proof relies on the proof presented by Bach and Moulines (2013) and is done separately for the noise process and for the noise-free process that depends only on the initial condition. The bounds may then be added (see the discussion in Appendix C.1).

Noise-free process. As in section A.3 from Bach and Moulines (2013), we assume here that the additive noise $(\xi_k^{\mathrm{add}})_{k \in \mathbb{N}^*}$ is uniformly equal to zero and that $\gamma(R_F^2 + \mathcal{M}_2) \leq 1$. Using Definitions 2 and 4, we thus have for any k in $\mathbb{N}^*$ that $\eta_k = \eta_{k-1} - \gamma H_F \eta_{k-1} + \gamma \xi_k^{\mathrm{mult}}(\eta_{k-1})$, it flows:

$$\begin{aligned}
\mathbb{E}[\|\eta_k\|^2] &= \mathbb{E}[\|\eta_{k-1}\|^2] - 2\gamma \mathbb{E}[\langle \eta_{k-1}, H_F \eta_{k-1} \rangle] + \gamma^2 \mathbb{E}[\|H_F \eta_{k-1} - \xi_k^{\mathrm{mult}}(\eta_{k-1})\|^2] \\
&= \mathbb{E}[\|\eta_{k-1}\|^2] - 2\gamma \mathbb{E}[\langle \eta_{k-1}, H_F \eta_{k-1} \rangle] + \gamma^2 \mathbb{E}[\|H_F \eta_{k-1}\|^2] + \gamma^2 \mathbb{E}[\|\xi_k^{\mathrm{mult}}(\eta_{k-1})\|^2] .
\end{aligned}$$

Considering that $H_F \preccurlyeq \mathrm{Tr}\left(H_F\right) \mathrm{I}_d \preccurlyeq R_F^2 \mathrm{I}_d$ and using Assumption 3, we obtain:

$$\mathbb{E}[\|\eta_k\|^2] \leq \mathbb{E}[\|\eta_{k-1}\|^2] - 2\gamma \mathbb{E}[\|H_F^{1/2} \eta_{k-1}\|^2] + \gamma^2 (R_F^2 + \mathcal{M}_2) \mathbb{E}[\|H_F^{1/2} \eta_{k-1}\|^2] .$$

Because the step-size γ is s.t. $\gamma(R_F^2 + \mathcal{M}_2) \leq 1$, we recover that in the absence of noise, we have:

$$\mathbb{E}[\|H_F^{1/2} \overline{\eta}_{K-1}\|^2] \leq \frac{\|\eta_0\|^2}{\gamma K} . \tag{S21}$$

Noise process. Now, all the following results comes from Appendix C.2 where we assume that $\eta_0 = w_0 - w_* = 0$, we start using Minkowski's inequality S6:

$$\mathbb{E}\left[\|H_F^{1/2} \overline{\eta}_{K-1}\|^2\right]^{1/2} \leq \mathbb{E}\left[\|H_F^{1/2} \sum_{i=0}^{r} \overline{\eta}_{K-1}^i\|^2\right]^{1/2} + \mathbb{E}\left[\|H_F^{1/2} (\overline{\eta}_{K-1} - \sum_{i=0}^{r} \overline{\eta}_{K-1}^i)\|^2\right]^{1/2} . \tag{S22}$$

First term.

Let $r \in \mathbb{N}$, again using Minkowski's inequality S6, we have

$$\begin{aligned}
\mathbb{E}[\|H_F^{1/2} \sum_{i=0}^{r} \overline{\eta}_{K-1}^i\|^2]^{1/2} &\leq \sum_{i=0}^{r} \mathbb{E}[\|H_F^{1/2} \overline{\eta}_{K-1}^i\|^2]^{1/2} \\
&= \mathbb{E}[\|H_F^{1/2} \overline{\eta}_{K-1}^0\|^2]^{1/2} + \sum_{i=1}^{r} \mathbb{E}[\|H_F^{1/2} \overline{\eta}_{K-1}^i\|^2]^{1/2} .
\end{aligned} \tag{S23}$$

By Equation (S20), we have $\eta_k^0 = (\mathrm{I}_d - \gamma H_F)\eta_{k-1}^0 + \gamma \xi_k^{\mathrm{add}}$, hence to bound the first term, we have to apply Lemma 2 from Bach and Moulines (2013) to the sequence $(\eta_{k-1}^0)_{k\in\mathbb{N}^*}$ and we obtain

$$\mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}^0\|^2] \leq \mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}} H_F^{-1}\right)/K\,. \tag{S24}$$

Let i in $\{1, \cdots, r\}$, to bound the second term, we have to apply Lemma 2 from Bach and Moulines (2013) to the sequence $(\eta_{k-1}^i)_{k\in\mathbb{N}^*}$. To do so, we bound the covariance of the noise which is here equal to $\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^{i-1}\right)$ (by definition of η_{k-1}^i, see Equation (S20)).
Because the multiplicative noise is linear, using Assumption 3, there exists a matrix Ξ_k in $\mathbb{R}^{d\times d}$ s.t. $\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^{i-1}\right) = \Xi_k \eta_{k-1}^{i-1}$. It follows that taking the expectation w.r.t to the σ-algebra $\sigma(\Xi_k)$, and because η_{k-1}^{i-1} is independent of it, using Lemma S44, we have:

$$\mathbb{E}\left[\eta_{k-1}^{i-1} \otimes \eta_{k-1}^{i-1} \mid \sigma(\Xi_k)\right] = \mathbb{E}\left[\eta_{k-1}^{i-1} \otimes \eta_{k-1}^{i-1}\right] \preccurlyeq \gamma^i Ш_{\mathrm{add}} Ш_{\mathrm{mult}}^{i-1} \mathrm{I}_d\,.$$

Thus, the noise $\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^{i-1}\right)$ is such that:

$$\mathbb{E}[\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^{i-1}\right) \otimes \xi_k^{\mathrm{mult}}\left(\eta_{k-1}^{i-1}\right) \mid \sigma(\Xi_k)] = \Xi_k \mathbb{E}\left[\eta_{k-1}^{i-1} \otimes \eta_{k-1}^{i-1}\right] \Xi_k^\top \preccurlyeq \gamma^i Ш_{\mathrm{add}} Ш_{\mathrm{mult}}^{i-1} \Xi_k \Xi_k^\top\,.$$

Taking full expectation, we furthermore consider Assumption 4.2 which gives that: $\mathbb{E}\left[\Xi_i \Xi_i^\top\right] \preccurlyeq Ш_{\mathrm{mult}} H_F$, hence:

$$\mathbb{E}\left[\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^{i-1}\right) \otimes \xi_k^{\mathrm{mult}}\left(\eta_{k-1}^{i-1}\right)\right] \leq \gamma^i Ш_{\mathrm{add}} Ш_{\mathrm{mult}}^{i} H_F\,. \tag{S25}$$

Using Lemma 2 from Bach and Moulines (2013) results to:

$$\sum_{i=1}^r \mathbb{E}[\|H_F^{1/2}\overline{\eta}_{K-1}^i\|^2]^{1/2} \leq \sum_{i=1}^r \gamma^i Ш_{\mathrm{add}} Ш_{\mathrm{mult}}^{i} \mathrm{Tr}\left(H_F H_F^{-1}\right)/K\,. \tag{S26}$$

In the end, we obtain from Equation (S23):

$$\begin{aligned}
\mathbb{E}[\|H_F^{1/2}\sum_{i=0}^r \overline{\eta}_{K-1}^i\|^2]^{1/2} &\leq \frac{\sqrt{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}} H_F^{-1}\right)}}{\sqrt{K}} + \frac{\sqrt{dШ_{\mathrm{add}}}}{\sqrt{K}} \sum_{i=1}^r \gamma^{i/2} Ш_{\mathrm{mult}}^{i/2} \\
&\leq \frac{\sqrt{\mathrm{Tr}\left(\mathfrak{C}_{\mathrm{ania}} H_F^{-1}\right)}}{\sqrt{K}} + \frac{\sqrt{\gamma d Ш_{\mathrm{add}} Ш_{\mathrm{mult}}}\left(1 - (\gamma Ш_{\mathrm{mult}})^{r/2}\right)}{\sqrt{K}\left(1 - \sqrt{\gamma Ш_{\mathrm{mult}}}\right)}\,.
\end{aligned}$$

Second term.
If $\gamma(R_F^2 + \mathcal{M}_2) \leq 1$, Lemma S43 gives:

$$\mathbb{E}\left\langle \overline{\eta}_{K-1} - \sum_{i=0}^r \overline{\eta}_{K-1}^i, H(\overline{\eta}_{K-1} - \sum_{i=0}^r \overline{\eta}_{K-1}^i)\right\rangle \leq \frac{\gamma}{(1-\gamma(R_F^2 + \mathcal{M}_2))K} \sum_{k=1}^K \mathbb{E}\left[\|\xi_k^{\mathrm{mult}}\left(\eta_{k-1}^r\right)\|^2\right]\,. \tag{S27}$$

Furthermore, $\left\|\xi_k^{\text{mult}}\left(\eta_{k-1}^r\right)\right\|^2 = \operatorname{Tr}\left(\xi_k^{\text{mult}}\left(\eta_{k-1}^r\right)^{\otimes 2}\right)$, by reusing what has been written in the previous paragraph (Equation (S25)), we obtain:

$$\begin{aligned}\|\xi_k^{\text{mult}}\left(\eta_{k-1}^r\right)\|^2 &\leq \gamma^{r+1}\text{Ш}_{\text{add}}\text{Ш}_{\text{mult}}^{r+1}\operatorname{Tr}\left(H_F\right)\\ &\leq \gamma^{r+1}\text{Ш}_{\text{add}}\text{Ш}_{\text{mult}}^{r+1}R_F^2 \qquad \text{(Definition 2).}\end{aligned}$$

It follows that we have:

$$\mathbb{E}\left\langle \overline{\eta}_{K-1} - \sum_{i=0}^{r}\overline{\eta}_{K-1}^i, H(\overline{\eta}_{K-1} - \sum_{i=0}^{r}\overline{\eta}_{K-1}^i)\right\rangle \leq \frac{\gamma^{r+2}\text{Ш}_{\text{add}}\text{Ш}_{\text{mult}}^{r+1}R_F^2}{(1-\gamma(R_F^2+\mathcal{M}_2))}. \tag{S28}$$

Putting things together. In the end, from the Minkowski decomposition done in Equation (S22), we combine the two terms and it leads to:

$$\begin{aligned}\mathbb{E}\left[\left\langle\overline{\eta}_{K-1}, H_F\overline{\eta}_{K-1}\right\rangle\right]^{1/2} \leq& \left(\frac{\gamma^{r+2}\text{Ш}_{\text{add}}\text{Ш}_{\text{mult}}^{r+1}R_F^2}{(1-\gamma(R_F^2+\mathcal{M}_2))}\right)^{1/2} + \frac{\sqrt{\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}}H_F^{-1}\right)}}{\sqrt{K}}\\ &+ \frac{\sqrt{\gamma d\text{Ш}_{\text{add}}\text{Ш}_{\text{mult}}}\left(1-(\gamma\text{Ш}_{\text{mult}})^{r/2}\right)}{\sqrt{K}\left(1-\sqrt{\gamma\text{Ш}_{\text{mult}}}\right)}.\end{aligned}$$

This implies that for any $\gamma\text{Ш}_{\text{mult}} \leq 1$, we obtain, by letting r tend to $+\infty$:

$$\mathbb{E}\left[\left\langle\overline{\eta}_{K-1}, H_F\overline{\eta}_{K-1}\right\rangle\right]^{1/2} \leq \frac{1}{\sqrt{K}}\left(\sqrt{\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}}H_F^{-1}\right)} + \frac{(\gamma d\text{Ш}_{\text{add}}\text{Ш}_{\text{mult}})^{1/2}}{1-\sqrt{\gamma\text{Ш}_{\text{mult}}}}\right). \tag{S29}$$

Final bound. We now take results derived from the part without noise, and the part with noise, to get:

$$\mathbb{E}[\langle\overline{\eta}_{K-1}, H_F\overline{\eta}_{K-1}\rangle]^{1/2} \leq \frac{1}{\sqrt{K}}\left(\frac{\|\eta_0\|}{\sqrt{\gamma}} + \sqrt{\operatorname{Tr}\left(\mathfrak{C}_{\text{ania}}H_F^{-1}\right)} + \frac{(\gamma d\text{Ш}_{\text{add}}\text{Ш}_{\text{mult}})^{1/2}}{1-\sqrt{\gamma\text{Ш}_{\text{mult}}}}\right),$$

which leads to the desired result considering that $4\gamma\text{Ш}_{\text{mult}} \leq 1$. ■

Appendix D. Validity of the assumptions made on the random fields

In this section, we verify that all the assumptions on the random fields done in Subsection 2.1 are fulfilled in the setting of compressed least-squares regression analyzed in Section 3. To do so, we first need to define the filtrations considered in this section.
For k in $\mathbb{N}^*$, we note u_k the noise that controls the randomization $\mathcal{C}_k(\cdot)$ at round k. In Section 2, we have denoted by $\mathcal{F}_k$ the σ-algebra generated by $(x_1, \varepsilon_1, u_1, \cdots, x_k, \varepsilon_k)$. In particular, w_k and $\overline{w}_k$ are $\mathcal{F}_k$-measurable. We also consider the following filtrations.

Definition S46 *We note $(\mathcal{G}_k)_{k\in\mathbb{N}}$ the filtration associated with the features noise, $(\mathcal{H}_k)_{k\in\mathbb{N}}$ the filtration associated with the output noise, and $(\mathcal{I}_k)_{k\in\mathbb{N}}$ the filtration associated with the* *stochastic gradient noise, which is the union of the two previous filtrations. Thus, we define* $\mathcal{F}_0 = \{\varnothing\}$ *and for* $k \in \mathbb{N}^*$*:*

$$\begin{aligned}
\mathcal{G}_k &= \sigma\left(\mathcal{F}_{k-1} \cup \{x_k\}\right) \\
\mathcal{H}_k &= \sigma\left(\mathcal{F}_{k-1} \cup \{\varepsilon_k\}\right) \\
\mathcal{I}_k &= \sigma\left(\mathcal{F}_{k-1} \cup \{x_k, \varepsilon_k\}\right) \\
\mathcal{F}_k &= \sigma\left(\mathcal{F}_{k-1} \cup \{x_k, \varepsilon_k, u_k\}\right) \,.
\end{aligned}$$

Note that there are two filtrations $\mathcal{G}$ *and* $\mathcal{H}$ *for the two independent noises that are both involved to compute the stochastic gradient. This will help us to compute the scalar product of these two quantities.*

We start by providing a bound on the distance between two compressions, this lemma will be used to prove Property S50.

Lemma S47 *For any compressor* $\mathcal{C}$ *in* $\mathbb{C}$ *verifying Lemma 16, for all* x, y *in* $\mathbb{R}^d$*, we have:*

$$\mathbb{E}[\|\mathcal{C}(x) - \mathcal{C}(y)\|^2] \leq 2(\omega+1)\left\|x\right\|^2 + 2(\omega+1)\left\|y\right\|^2 \,.$$

Proof Let a compressor $\mathcal{C}$ in $\mathbb{C}$ and x, y in $\mathbb{R}^d$, using Inequality 1, we have that:

$$\left\|\mathcal{C}(x) - \mathcal{C}(y)\right\|^2 \leq 2\left\|\mathcal{C}(x)\right\|^2 + 2\left\|\mathcal{C}(y)\right\|^2 \,.$$

Taking expectation and using Lemma 16 allows to conclude:

$$\mathbb{E}\left[\left\|\mathcal{C}(x) - \mathcal{C}(y)\right\|^2\right] \leq 2(\omega+1)\left\|x\right\|^2 + 2(\omega+1)\left\|y\right\|^2 \,.$$

■

Now we prove that all the assumptions done in Section 2 are correct.

Property S48 (Validity of the setting presented in Definition 2) *Consider the Algorithm 2 in the context of Model 2, we have that the setting presented in Definition 2 is verified.*

Proof From Algorithm 2, we have for any k in $\mathbb{N}^*$ and any w in $\mathbb{R}^d$ $\xi_k(w - w_*) = \nabla F(w) - \mathcal{C}_k(g_k(w))$. Because $(g_k)_{k \in \mathbb{N}^*}$ and $(\mathcal{C}_k)_{k \in \mathbb{N}^*}$ are by definition two sequences of i.i.d. random fields (Algorithm 2), it follows that their composition is also i.i.d., *hence* $(\xi_k)_{k \in \mathbb{N}^*}$ *is a sequence of i.i.d. random fields.*
Taking expectation w.r.t. the σ-algebra $\mathcal{I}_k$, we have $\mathbb{E}\left[\mathcal{C}_k(g_k(w)) \mid \mathcal{I}_k\right] = g_k(w)$ (Lemma 16), next with the σ-algebra $\mathcal{F}_{k-1}$, we have $\mathbb{E}\left[g_k(w) \mid \mathcal{F}_{k-1}\right] = \nabla F(w)$ (Equation 2). *Hence, the random fields are zero-centered.*
From Model 2, we have for any k in $\mathbb{N}^*$ and any w in $\mathbb{R}^d$ that:

$$\begin{aligned}
F(w) &= \frac{1}{2}\mathbb{E}\left[(\langle x_k, w\rangle - y_k)^2\right] = \frac{1}{2}\mathbb{E}\left[(w - w_*)^\top (x_k \otimes x_k)(w - w_*) - 2\varepsilon_k \left\langle x_k, w - w_*\right\rangle + \varepsilon_k^2\right] \\
&= \frac{1}{2}((w - w_*)^\top H (w - w_*) + \sigma^2)\,,
\end{aligned}$$

hence F is quadratic with Hessian equal to H whose trace is equal to R^2. ■

Property S49 (Validity of Assumption 1) *Considering Algorithm 2 under the setting of Model 2 with Lemma 16, for any iteration k in $\mathbb{N}^*$, the second moment of the additive noise ξ_k^{add} can be bounded by $(\omega+1)R^2\sigma^2$, i.e., Assumption 1 is verified.*

Proof Let k in $\mathbb{N}^*$. Because we consider Algorithm 2, with Definitions 2 and 4, we first have $\xi_k^{\mathrm{add}} = -\mathcal{C}_k(g_{k,*})$, then with Lemma 16 we obtain $\mathbb{E}[\|\mathcal{C}_k(g_{k,*})\|^2 \mid \mathcal{I}_k] \leq (\omega+1)\|g_{k,*}\|^2$. Next, we first have from Model 2 and Equation (2) that $g_{k,*} = \varepsilon_k x_k$, secondly because $\left((\varepsilon_k)_{k\in[K]}\right)$ is independent from $\left((x_k)_{k\in[K]}\right)$ (Model 2), we have that $\mathbb{E}[\|\varepsilon_k x_k\|^2] \leq \sigma^2 R^2$, hence it results to:

$$\mathbb{E}[\|\xi_k^{\mathrm{add}}\|^2 \mid \mathcal{F}_{k-1}] = \mathbb{E}[\|\xi_k^{\mathrm{add}}\|^2] \leq (\omega+1)\sigma^2 R^2 \,.$$

■

Property S50 (Validity of Assumption 2.1) *Considering Algorithm 2, under the setting of Model 2 with Lemma 16, for any iteration k in $\mathbb{N}^*$, the second moment of the multiplicative noise $\xi_k^{\mathrm{mult}}(w)$ can be bounded for any w in $\mathbb{R}^d$ by $2(\omega+1)R^2\left\|H^{1/2}(w-w_*)\right\|^2 + 4(\omega+1)\sigma^2R^2$, i.e., Assumption 2.1 is verified.*

Proof Let k in $\mathbb{N}^*$, we note $\eta = w - w_*$. First, because we consider Algorithm 2, with Definitions 2 and 4, we have $\xi_k(\eta) = \nabla F(w) - \mathcal{C}_k(g_k(w))$ and $\xi_k^{\mathrm{add}} = -\mathcal{C}_k(g_{k,*})$, hence:

$$\xi_k^{\mathrm{mult}}(\eta) = \xi_k(\eta) - \xi_k^{\mathrm{add}} = \nabla F(w) - \mathcal{C}_k(g_k(w)) + \mathcal{C}_k(g_{k,*})\,,$$

thus developing the squared-norm of $\xi_k^{\mathrm{mult}}(\eta)$ gives:

$$\|\xi_k^{\mathrm{mult}}(\eta)\|^2 = \|\nabla F(w)\|^2 + 2\left\langle \nabla F(w), \mathcal{C}_k(g_{k,*}) - \mathcal{C}_k(g_k(w))\right\rangle + \|\mathcal{C}_k(g_{k,*}) - \mathcal{C}_k(g_k(w))\|^2 \,.$$

On the first side we have $\mathbb{E}\left[\mathbb{E}\left[\mathcal{C}_k(g_{k,*}) - \mathcal{C}_k(g_k(w)) \mid \mathcal{I}_k\right] \mid \mathcal{F}_{k-1}\right] = -\nabla F(w_{k-1})$. On the second side, we use Lemma S47; this allows us to write:

$$\mathbb{E}\left[\|\mathcal{C}_k(g_{k,*}) - \mathcal{C}_k(g_k(w))\|^2 \,\middle|\, \mathcal{I}_k\right] \leq 2(\omega+1)\|g_k(w)\|^2 + 2(\omega+1)\|g_{k,*}\|^2 \,.$$

Note that this bound is far from being optimal when $g_k(w) = g_{k,*}$ or if $\mathcal{C}$ is the identity. Next, we decompose as follows:

$$\begin{aligned}\mathbb{E}\left[\|\mathcal{C}_k(g_{k,*}) - \mathcal{C}_k(g_k(w))\|^2 \,\middle|\, \mathcal{I}_k\right] &\leq 2(\omega+1)\|g_k(w) - g_{k,*}\|^2 \\ &\quad + 4(\omega+1)\left\langle g_k(w) - g_{k,*}, g_{k,*}\right\rangle + 4(\omega+1)\|g_{k,*}\|^2 \,.\end{aligned}$$

Taking expectation w.r.t. the σ-algebra $\mathcal{G}_k$, recalling that $g_k(w) - g_{k,*}$ is $\mathcal{G}_k$-measurable (Definition S46) and considering Model 2 allows to write:

$$\begin{aligned}\mathbb{E}\left[\|\mathcal{C}_k(g_{k,*}) - \mathcal{C}_k(g_k(w)\|^2 \,\middle|\, \mathcal{G}_k\right] &\leq 2(\omega+1)\|g_{k,*} - g_k(w)\|^2 + 4(\omega+1)\sigma^2R^2 \\ &\leq 2(\omega+1)\|(x_k\otimes x_k)\eta_{k-1}\|^2 + 4(\omega+1)\sigma^2R^2\,,\end{aligned}$$

and now taking expectation w.r.t the σ-algebra $\mathcal{F}_{k-1}$ concludes the proof:

$$\mathbb{E}[\|\mathcal{C}_k(g_{k,*}) - \mathcal{C}_k(g_k(w))\|^2 \ \mid \mathcal{F}_{k-1}] \leq 2(\omega+1)R^2\|H^{1/2}(w_k - w_*)\|^2 + 4(\omega+1)\sigma^2 R^2\,.$$

■

Property S51 (Validity of Assumption 2.2) *Considering Algorithm 2, under the setting of Model 2 with Lemma 16, for any iteration k in $\mathbb{N}^*$, the second moment of the multiplicative noise $\xi_k^{\mathrm{mult}}(w)$ can be bounded for any w in $\mathbb{R}^d$ by $\Omega R^2\sigma\|H^{1/2}(w-w_*)\| + 3(\omega+1)R^2\|H^{1/2}(w-w_*)\|^2$, i.e. Assumption 2.2 is verified.*

Proof Let k in $\mathbb{N}^*$, we note $\eta = w - w_*$. Because we consider Algorithm 2, with Definitions 2 and 4, we have the following decomposition:

$$\xi_k^{\mathrm{mult}}(\eta) = \|\nabla F(w)\|^2 + 2\left\langle \nabla F(w), \mathcal{C}_k(g_{k,*}) - \mathcal{C}_k(g_k(w))\right\rangle + \|\mathcal{C}_k(g_{k,*}) - \mathcal{C}_k(g_k(w))\|^2\ .$$

We take expectation w.r.t. the σ-algebra $\mathcal{I}_k$ and use Item L.2 of Lemma 16:

$$\begin{aligned}\mathbb{E}\left[\xi_k^{\mathrm{mult}}(\eta) \ \Big|\ \mathcal{I}_k\right] \leq{}& \|\nabla F(w)\|^2 + 2\left\langle \nabla F(w), g_{k,*} - g_k(w)\right\rangle \\ &+ \Omega\min(\|g_{k,*}\|, \|g_k(w)\|)\|g_{k,*} - g_k(w)\| + 3(\omega+1)\|g_{k,*} - g_k(w)\|^2\,.\end{aligned}$$

Then, we have $\min(\|g_{k,*}\|, \|g_k(w)\|)\|g_{k,*} - g_k(w)\| \leq \|g_{k,*}\|\|g_{k,*} - g_k(w)\|$, taking expectation conditionally to the σ-algebra $\mathcal{F}_{k-1}$, applying the Cauchy-Schwarz's Equation (S8) and considering Model 2, we have:

$$\begin{aligned}\mathbb{E}[\|g_{k,*}\|\|g_{k,*} - g_k(w)\| \mid \mathcal{F}_{k-1}]^2 &\leq \mathbb{E}[\|g_{k,*}\|^2 \mid \mathcal{F}_{k-1}]\mathbb{E}[\|g_{k,*} - g_k(w)\|^2 \mid \mathcal{F}_{k-1}] \\ &\leq \sigma^2 R^4\|H^{1/2}(w-w_*)\|^2\,.\end{aligned}$$

Therefore, we can conclude:

$$\mathbb{E}\left[\xi_k^{\mathrm{mult}}(\eta) \ \Big|\ \mathcal{F}_{k-1}\right] \leq -\|\nabla F(w)\|^2 + \sigma R^2\Omega\|H^{1/2}(w-w_*)\| + 3(\omega+1)R^2\|H^{1/2}(w-w_*)\|^2\,.$$

■

Property S52 (Validity of Assumption 3) *Considering Algorithm 2, under the setting of Model 2 with Lemma 16, if the compressor $\mathcal{C}$ is linear, then for any iteration k in $\mathbb{N}^*$, the multiplicative noise ξ_k^{mult} is linear, thus there exist a matrix Ξ_k in $\mathbb{R}^{d\times d}$ such that for any w in $\mathbb{R}^d$, $\xi_k^{\mathrm{mult}}(w) = \Xi_k w$. Furthermore, the second moment of the multiplicative noise can be bounded for any w in $\mathbb{R}^d$ by $(\omega+1)R^2\left\|H^{1/2}(w-w_*)\right\|^2$, hence Assumption 3 is verified.*

Proof Let k in $\mathbb{N}^*$, we note $\eta = w - w_*$. First, because we consider Algorithm 2, with Definitions 2 and 4, we have $\xi_k(\eta) = \nabla F(w) - \mathcal{C}_k(g_k(w))$ and $\xi_k^{\mathrm{add}} = -\mathcal{C}_k(g_{k,*})$, hence:

$$\xi_k^{\mathrm{mult}}(\eta) = \xi_k(\eta) - \xi_k^{\mathrm{add}} = \nabla F(w) - \mathcal{C}_k(g_k(w)) + \mathcal{C}_k(g_{k,*})\,.$$

Because the random mechanism $\mathcal{C}_k$ is linear, there exists a random matrix Π_k in $\mathbb{R}^{d\times d}$ such that for any z in $\mathbb{R}^d$, we have $\mathcal{C}_k(z) = \Pi_k z$, it follows that:

$$\xi_k^{\mathrm{mult}}(\eta) = \nabla F(w) + \mathcal{C}_k(g_{k,*} - g_k(w)) = (H - \Pi_k(x_k \otimes x_k))\eta\,.$$

Hence, the first part of Assumption 3 is verified with $\Xi_k = H - \Pi_k(x_k \otimes x_k)$. Now, we compute the second moment of the multiplicative noise. We start by developing its squared norm:

$$\|\xi_k^{\mathrm{mult}}(\eta)\|^2 = \|\nabla F(w)\|^2 + 2\left\langle \nabla F(w), \mathcal{C}_k(g_{k,*} - g_k(w))\right\rangle + \|\mathcal{C}_k(g_{k,*} - g_k(w))\|^2\,.$$

Taking expectation conditionally to the σ-algebra $\mathcal{I}_k$, and using Lemma 16 gives:

$$\mathbb{E}\left[\|\xi_k^{\mathrm{mult}}(\eta))\|^2 \;\middle|\; \mathcal{I}_k\right] = \|\nabla F(w)\|^2 + 2\left\langle \nabla F(w), g_{k,*} - g_k(w)\right\rangle + (\omega+1)\|g_{k,*} - g_k(w)\|^2\,.$$

Finally, with σ-algebra $\mathcal{F}_{k-1}$ and considering Model 2 we have:

$$\mathbb{E}\left[\|\xi_k^{\mathrm{mult}}(\eta)\|^2 \;\middle|\; \mathcal{F}_{k-1}\right] = -\|\nabla F(w)\|^2 + (\omega+1)R^2\|H^{1/2}(w-w_*)\|^2\,,$$

which allows to conclude. ■

Property S53 (Validity of Assumption 4) *Considering Algorithm 2 under the setting of Model 2 with Remark 1 and Lemma 16, if the compressor $\mathcal{C}$ is linear, then for any k in $\mathbb{N}^*$, there exists a constant $Ш_H > 0$ s.t. $\mathfrak{C}_{\mathrm{ania}} \preccurlyeq \sigma^2 Ш_H H_F$ and $\mathbb{E}\left[\Xi_k \Xi_k^\top\right] \preccurlyeq R^2 Ш_H H$; Assumption 4 is thus verified.*

Proof

Let k in $\mathbb{N}^*$, we note $\eta = w - w_*$. We first need to compute $Ш_H$ in $\mathbb{R}^d$ for each compressor $\mathcal{C}$ in $\{\mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{sq}}, \mathcal{C}_{\mathrm{rd1}}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_\Phi, \mathcal{C}_{\mathrm{PP}}\}$, it comes from Proposition 21 which results having a constant $Ш_H$ s.t.:

$$\mathfrak{C}(\mathcal{C}, p_H) = \mathbb{E}_{E\sim p_H}[\mathcal{C}(E)^{\otimes 2}] \preccurlyeq Ш_H H\,. \tag{S30}$$

Indeed, $\mathrm{Diag}\,(H)$ car be bounded by $\mathrm{Tr}\,(H)\,\mathrm{I}_d$, and then I_d by $\mu^{-1}H$. This constant $Ш_H$ can be computed from Proposition 21 for any compressor:

Compressor	$\mathcal{C}_{\mathrm{rd}h}$	$\mathcal{C}_{\mathrm{s}}$	$\mathcal{C}_{\mathrm{PP}}$	$\mathcal{C}_\Phi$
$Ш_H$	$\frac{h-1}{p(d-1)} + (1 - \frac{h-1}{d-1})\frac{\tau}{p}$	$1 + \frac{(1-p)\tau}{p}$	$\frac{1}{p}$	$\frac{\alpha-\beta}{p} + \frac{\beta\tau}{p}$
$Ш_H$ (if H diagonal)	$\frac{1}{p}$	$\frac{1}{p}$	$\frac{1}{p}$	$\frac{\alpha-\beta}{p} + \frac{\beta\tau}{p}$

Where $p = h/d$, $\tau = \mathrm{Tr}\,(H)\,/\mu$, and for sketching $\alpha = \frac{h+2}{d+2}$ and $\beta = \frac{d-h}{(d-1)(d+2)}$.
We now show that the two inequalities given in Assumption 4 are valid.
First inequality.
By Definition 6, we have $\mathfrak{C}_{\mathrm{ania}} = \mathbb{E}\left[\xi_k^{\mathrm{add}} \otimes \xi_k^{\mathrm{add}}\right] = \mathbb{E}\left[\mathcal{C}_k(\varepsilon_k x_k)^{\otimes 2}\right]$, because $\left((\varepsilon_k)_{k\in[K]}\right)$ is independent from $\left((x_k)_{k\in[K]}\right)$ (Model 2) and using compressor linearity and Equation (S30), it gives: $\mathfrak{C}_{\mathrm{ania}} = \sigma^2\mathbb{E}\left[\mathcal{C}_k(x_k)^{\otimes 2}\right] = \sigma^2\mathfrak{C}(\mathcal{C}, p_H) \preccurlyeq \sigma^2 Ш_H H\,.$

Second inequality.
Using Property S53, because the compressor $\mathcal{C}$ is linear, there exists two matrices Π_k, Ξ_k in $\mathbb{R}^{d\times d}$ s.t. for any z in $\mathbb{R}^d$, we have $\mathcal{C}_k(z) = \Pi_k z$ and $\xi_k^{\mathrm{mult}}(z) = \Xi_k z$, which gives that $\Xi_k = H - \Pi_k(x_k \otimes x_k)$. It follows that:

$$\Xi_k \Xi_k^\top = HH^\top - H\Pi_k(x_k \otimes x_k) - \Pi_k(x_k \otimes x_k)H + \Pi_k(x_k \otimes x_k)(x_k \otimes x_k)\Pi_k^\top .$$

Given that the compression is unbiased (Lemma 16) we have $\mathbb{E}\left[\Pi_k \mid \mathcal{I}_k\right] = \mathrm{I}_d$, hence:

$$\mathbb{E}\left[\Xi_k \Xi_k^\top \;\middle|\; \mathcal{I}_k\right] = HH^\top - H(x_k \otimes x_k) - (x_k \otimes x_k)H + \mathbb{E}\left[\Pi_k(x_k \otimes x_k)(x_k \otimes x_k)\Pi_k^\top \;\middle|\; \mathcal{I}_k\right] ,$$

and now taking expectation w.r.t the σ-algebra $\mathcal{F}_{k-1}$:

$$\mathbb{E}\left[\Xi_k \Xi_k^\top \;\middle|\; \mathcal{F}_{k-1}\right] = -HH^\top + \mathbb{E}\left[\Pi_k(x_k \otimes x_k)(x_k \otimes x_k)\Pi_k^\top \;\middle|\; \mathcal{F}_{k-1}\right] .$$

In the end, we have that $\mathbb{E}\left[\Xi_k \Xi_k^\top \mid \mathcal{F}_{k-1}\right] \preccurlyeq \mathbb{E}\left[\Pi_k(x_k \otimes x_k)(x_k \otimes x_k)\Pi_k^\top \mid \mathcal{F}_{k-1}\right]$, and if we consider that the second moment of the features $(x_k)_{k\in\mathbb{N}^*}$ is almost surely bounded (Remark 1), we obtain:

$$\mathbb{E}\left[\Xi_k \Xi_k^\top \;\middle|\; \mathcal{F}_{k-1}\right] \preccurlyeq R^2 \mathbb{E}\left[\Pi_k(x_k \otimes x_k)\Pi_k^\top \;\middle|\; \mathcal{F}_{k-1}\right] \preccurlyeq R^2 \mathbb{E}\left[\mathcal{C}_k(x_k)^{\otimes 2} \mid \mathcal{F}_{k-1}\right] . \tag{S31}$$

Thus, using Equation (S30), we can state that $\mathbb{E}\left[\Xi_k \Xi_k^\top \mid \mathcal{F}_{k-1}\right] \preccurlyeq R^2 \text{Ш}_H H$, which concludes the second part of the verification of Assumption 4.

■

Appendix E. Compression operators

In this Section, we provide additional details about compression operators. First, we prove in Appendix E.1 that Lemma 16 hold and compute the compressor's covariance given in Proposition 21. The specific computations for sketching are given separately in Appendix E.2 because they are more complex. Third, it allows to prove Propositions 24 and 25 in Appendix E.3. And finally, in Appendix E.4, we plot the covariance matrix induced by quantization and sparsification for `quantum` and `cifar-10`.

E.1 Computation of the variance and covariance of the compression operators

In this Subsection, we first prove Lemma 16. Item L.1 is frequently established in the literature and corresponds to the worst-case assumption, see the introduction for references. On the other hand, Item L.2 is the Hölder-type bound, which is not used in the literature up to our knowledge. Next, we compute the compressors' covariances that have been given in Proposition 21.

Lemma S54 *For any compressor $\mathcal{C} \in \{\mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{sq}}, \mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_\Phi, \mathcal{C}_{\mathrm{PP}}\}$, there exists constants $\omega, \Omega \in \mathbb{R}_+^*$, such that the random operator $\mathcal{C}$ satisfies the following properties for all $z, z' \in \mathbb{R}^d$.*

L.1: $\mathbb{E}[\mathcal{C}(z)] = z$ *and* $\mathbb{E}[\|\mathcal{C}(z)-z\|^2] \leq \omega\|z\|^2$ *(unbiasedness and variance relatively bounded),*
L.2: $\mathbb{E}[\|\mathcal{C}(z) - \mathcal{C}(z')\|^2] \leq \Omega \min(\|z\|, \|z'\|)\|z - z'\| + 3(\omega + 1)\|z - z'\|^2$ *(Hölder-type bound),*

with $\omega = \sqrt{d}$ *and* $\Omega = 12\sqrt{d}$ *(resp.* $\omega = (1-p)/p$ *and* $\Omega = 0$*) for* $\mathcal{C}_{\mathrm{q}}$ *and* $\mathcal{C}_{\mathrm{sq}}$ *(resp.* $\mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_{\mathrm{s}}$*,* $\mathcal{C}_{\Phi}, \mathcal{C}_{\mathrm{PP}}$*).*

Proof

Value of ω (Item L.1 of Lemma 16). For projection-based compressors, the proof is straightforward, for quantization-based, the proof can be found in Alistarh et al. (2017).

Value of Ω (Item L.2 of Lemma 16). For linear compressors, it is straightforward to obtain $\Omega = 0$.
For quantization, we take x, y in $\mathbb{R}^d$, we note $(u_i)_{i=1}^d$ the vector controlling the randomness of compression, and we write $\mathcal{C}_{\mathrm{q}}(x) - \mathcal{C}_{\mathrm{q}}(y) = A + B + C$, with:

1. $A := \|x\|\mathrm{sign}(x)\mathrm{Bern}(\frac{|x|}{\|x\|}) - \|x\|\mathrm{sign}(x)\mathrm{Bern}(\frac{|x|}{\|y\|})$
2. $B := \|x\|\mathrm{sign}(x)\mathrm{Bern}(\frac{|x|}{\|y\|}) - \|x\|\mathrm{sign}(y)\mathrm{Bern}(\frac{|y|}{\|y\|})$
3. $C := \|x\|\mathrm{sign}(y)\mathrm{Bern}(\frac{|y|}{\|y\|}) - \|y\|\mathrm{sign}(y)\mathrm{Bern}(\frac{|y|}{\|y\|})$.

We note $\|\cdot\|$ the 2-norm and $\|\cdot\|_1$ the 1-norm. By symmetry, we suppose that $\|y\|^2 \geq \|x\|^2$.
First term. We have $\|A\|^2 = \|x\|^2 \sum_{i=1}^d (\mathbb{1}_{u_i \leq \frac{|x_i|}{\|x\|}} - \mathbb{1}_{u_i \leq \frac{|x_i|}{\|y\|}})^2 = \|x\|^2 \sum_{i=1}^d \mathbb{1}^2_{\frac{|x_i|}{\|y\|} \leq u_i \leq \frac{|x_i|}{\|x\|}}$ because $\|y\|^2 \geq \|x\|^2$. Taking expectation, it gives $\mathbb{E}[\|A\|^2] = \|x\|^2 \sum_{i=1}^d \frac{|x_i|}{\|x\|} - \frac{|x_i|}{\|y\|} = \|x\|^2 \|x\|_1 \frac{\|y\| - \|x\|}{\|y\|\|x\|}$. Now with triangular inequality, we have:

$$\mathbb{E}[\|A\|^2] \leq \frac{\|x\|}{\|y\|}\|x\|_1\|y - x\| \leq \|x\|_1\|y - x\| \leq \sqrt{d}\|x\|\|y - x\|\,,$$

and by symmetry $\mathbb{E}[\|A\|^2] \leq \sqrt{d}\min(\|x\|, \|y\|)\|y - x\|$.
Second term.
We have $\|B\|^2 = \|x\|^2 \sum_{i=1}^d (\mathrm{sign}(x_i)\mathbb{1}_{u_i \leq \frac{|x_i|}{\|y\|}} - \mathrm{sign}(y_i)\mathbb{1}_{u_i \leq \frac{|y_i|}{\|y\|}})^2$. Let i in $[d]$, if $\mathrm{sign}(x_i) = \mathrm{sign}(y_i)$, then:

$$\mathbb{E}\left[\|B\|^2\right] = \|x\|^2 \sum_{i=1}^d \mathbb{E}\left[\mathbb{1}^2_{\frac{\min(|x_i|,|y_i|)}{\|y\|} \leq u_i \leq \frac{\max(|x_i|,|y_i|)}{\|y\|}}\right] = \frac{\|x\|^2}{\|y\|}\sum_{i=1}^d |y_i - x_i| \leq \|x\|\|x - y\|_1\,.$$

If $\mathrm{sign}(x_i) \neq \mathrm{sign}(y_i)$, developping $(\mathrm{sign}(x_i)\mathbb{1}_{u_i \leq \frac{|x_i|}{\|y\|}} - \mathrm{sign}(y_i)\mathbb{1}_{u_i \leq \frac{|y_i|}{\|y\|}})^2$, we have:

$$\begin{aligned}
\mathbb{E}\left[\|B\|^2\right] &= \|x\|^2 \sum_{i=1}^d \frac{|x_i|}{\|y\|} + \frac{|y_i|}{\|y\|} - 2\mathrm{sign}(x_i)\mathrm{sign}(y_i)\frac{\min(|x_i|, |y_i|)}{\|y\|} \\
&= \frac{\|x\|^2}{\|y\|}\sum_{i=1}^d \max(|x_i|, |y_i|) + 3\min(|x_i|, |y_i|)\,.
\end{aligned}$$

Next, we have $\max(|x_i|, |y_i|) + \min(|x_i|, |y_i|) = |x_i| + |y_i| \overset{\mathrm{sign}(x_i) \neq \mathrm{sign}(y_i)}{=} |x_i - y_i|$, which results to $\mathbb{E}\left[\|B\|^2\right] \leq 3\frac{\|x\|^2}{\|y\|}\sum_{i=1}^d |y_i - x_i| \leq 3\|x\|\|x - y\|_1 \leq 3\sqrt{d}\|x\|\|x - y\|$.

Third term. We have $\|C\|^2 = (\|x\| - \|y\|)^2 \sum_{i=1}^d \mathbb{1}^2_{u_i \leq \frac{|y_i|}{\|y\|}}$, taking expectation, it gives:

$$\mathbb{E}[\|C\|^2] = (\|x\| - \|y\|)^2 \sum_{i=1}^d \frac{|y_i|}{\|y\|} \leq \|x-y\|^2 \frac{\|y\|_1}{\|y\|} \leq \sqrt{d}\|x-y\|^2 \,.$$

Overall, using Inequality 1, we have:

$$\mathbb{E}[\|\mathcal{C}_{\mathrm{q}}(x) - \mathcal{C}_{\mathrm{q}}(y)\|^2] \leq 12\sqrt{d}\min(\|x\|, \|y\|)\|x-y\| + 3\sqrt{d}\|x-y\|^2 \,,$$

which allows to conclude as for 1-quantization, we have $\omega = \sqrt{d}$. ■

We now compute the compressors' covariance given in Proposition 21 and Corollary 22. However, sketching requires more involved computations, they are provided in Appendix E.2.

Proposition S55 (Structure of the compressor's covariance) *The following formulas of compressors' covariance hold:*

- $\mathfrak{C}(\mathcal{C}_\emptyset, p_M) = M$
- $\mathfrak{C}(\mathcal{C}_q, p_M) \preccurlyeq M + \sqrt{\mathrm{Tr}\,(M)}\sqrt{\mathrm{Diag}\,(M)} - \mathrm{Diag}\,(M)$
- $\mathfrak{C}(\mathcal{C}_s, p_M) = M + \frac{1-p}{p}\mathrm{Diag}\,(M)$
- $\mathfrak{C}(\mathcal{C}_\Phi, p_M) = \frac{1}{p}\left((\alpha-\beta)M + \beta\mathrm{Tr}\,(M)\,\mathrm{I}_d\right)$ *with* $\alpha = \frac{h+2}{d+2}$ *and* $\beta = \frac{d-h}{(d-1)(d+2)}$
- $\mathfrak{C}(\mathcal{C}_{\mathrm{rd}h}, p_M) = \frac{d(h-1)}{h(d-1)}M + \left(\frac{d}{h} - \frac{d(h-1)}{h(d-1)}\right)\mathrm{Diag}\,(M)$
- $\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) = \frac{1}{p}M$.

Proof
In this proof, we denote $\mathcal{F}$ the σ-field generated by the random sampling of $E \sim p_M \in \mathcal{P}_M$, and $\mathcal{G}$ the σ-field generated by the noise from the compression process. Let $E \sim p_M \in \mathcal{P}_M$.

Quantization. By definition, we have $\mathcal{C}_q(E) = \|E\|_2 \mathrm{sign}(E) \odot \chi$, with $\chi = \left(\mathrm{Bern}(\frac{|E_i|}{\|E\|_2})\right)_{i=1}^d$. It follows that $\mathcal{C}_q(E)^{\otimes 2} = \|E\|_2^2 \mathrm{sign}(E)^{\otimes 2} \odot \chi^{\otimes 2}$.
Because:

$$\mathbb{E}\left[\chi^{\otimes 2} \mid \mathcal{F}\right] = \begin{cases} \dfrac{|E_i|}{\|E\|_2} & \text{if} \quad i=j \\[2ex] \dfrac{|E_i|\,|E_j|}{\|E\|_2^2} & \text{else,} \end{cases}$$

and considering that $\mathrm{sign}(E)^{\otimes 2} = \begin{pmatrix} 1 & & \mathrm{sign}(E_i)\mathrm{sign}(E_j) \\ & \ddots & \\ \mathrm{sign}(E_i)\mathrm{sign}(E_j) & & 1 \end{pmatrix}$, we have:

$$\mathbb{E}\left[\mathcal{C}_q(E)^{\otimes 2} \mid \mathcal{F}\right] = \begin{cases} \|E\|_2\,|E_i| & \text{if} \quad i=j\,, \\[1ex] E_iE_j & \text{else.} \end{cases}$$

Taking the complete expectation gives:

$$\mathbb{E}\left[\mathcal{C}_q(E)^{\otimes 2}\right] = \begin{cases} \mathbb{E}\left[\|E\|_2\,|E_i|\right] & \text{if} \quad i=j \\[1ex] M_{ij} & \text{else.} \end{cases}$$

Changing the diagonal to make appear M, we obtain:

$$\mathbb{E}\left[\mathcal{C}_q(E)^{\otimes 2}\right] = M + \mathbb{E}\left[\|E\|_2 \mathrm{Diag}\left(|E_i|\right)_{i=1}^d\right] - \mathbb{E}\left[\mathrm{Diag}\left(E_i^2\right)_{i=1}^d\right] .$$

Furthermore, we first have that $\mathbb{E}\left[\mathrm{Diag}\left(E_i^2\right)_{i=1}^d\right] = \mathrm{Diag}\left(M\right)$ and secondly, by Cauchy-Schwarz Equation (S8) that:

$$\mathbb{E}\left[\|E\|_2 \mathrm{Diag}\left(|E_i|\right)_{i=1}^d\right]^2 \preccurlyeq \mathbb{E}\left[\|E\|_2^2\right] \mathbb{E}\left[\mathrm{Diag}\left(E_i^2\right)_{i=1}^d\right] = \mathrm{Tr}\left(M\right) \mathrm{Diag}\left(M\right) ,$$

which finally gives $\mathbb{E}\left[\mathcal{C}_q(E)^{\otimes 2}\right] \preccurlyeq M + \sqrt{\mathrm{Tr}\left(M\right)}\sqrt{\mathrm{Diag}\left(M\right)} - \mathrm{Diag}\left(M\right)$.

Sparsification. By definition, we have $\mathcal{C}_s(E) = \frac{1}{p} B \odot E \in \mathbb{R}^d$, with $B \sim (\mathrm{Bern}(p))_{i=1}^d$, thus $\mathcal{C}_s(E)^{\otimes 2} = \frac{1}{p^2} B^{\otimes 2} \odot E^{\otimes 2}$. Taking the expectation w.r.t. to the σ-filtration $\mathcal{F}$, we have $\mathbb{E}\left[\mathcal{C}_s(E)^{\otimes 2} \mid \mathcal{F}\right] = \frac{1}{p^2} P \odot E^{\otimes 2}$ with $P = \begin{pmatrix} p & & p^2 \\ & \ddots & \\ p^2 & & p \end{pmatrix}$, because for all i, j in $[\![1, d]\!]$, we have $\mathbb{E}\left[B_i^2 \mid \mathcal{F}\right] = p$ and $\mathbb{E}\left[B_i B_j \mid \mathcal{F}\right] = p^2$. This naturally gives: $\mathbb{E}\left[\mathcal{C}_s(E)^{\otimes 2}\right] = \frac{1}{p^2} P \odot M$.

Sketching. The proof is more complex and therefore is given separately, in Appendix E.2.3.

Rand-h. By definition, we have $\mathcal{C}_{\mathrm{rd}h}(E) := \frac{d}{h} B(S) \odot E$ with $S \sim \mathrm{Unif}(\mathcal{P}_h([d]))$ and $B(S)_i = \mathbb{1}_{i \in S}$, thus $\mathcal{C}_{\mathrm{rd}h}(E)^{\otimes 2} = \frac{1}{p^2} B^{\otimes 2} \odot E^{\otimes 2}$ $(p = h/d)$. We have that for any i, j in $\{1, \dots, d\}$, B_i and B_j are *not* independent and that $B_i \sim (\mathrm{Bern}(p))$, therefore we have that $\mathbb{E}[B_i^2] = p$ and that: $h^2 = \left(\sum_{i=1}^d B_i\right)^2 = \sum_{i=1}^d B_i^2 + \sum_{i \neq j} B_i B_j$. Taking expectation, it gives $h^2 = h + d(d-1)\mathbb{E}[B_i B_j]$ i.e. $\mathbb{E}[B_i B_j] = \frac{h(h-1)}{d(d-1)}$. Taking the expectation w.r.t. to the σ-filtration $\mathcal{F}$, we have :

$$\mathbb{E}\left[\mathcal{C}_{\mathrm{rd}h}(E)^{\otimes 2} \mid \mathcal{F}\right] = \frac{d(h-1)}{h(d-1)} E^{\otimes 2} + \left(\frac{d}{h} - \frac{d(h-1)}{h(d-1)}\right) \mathrm{Diag}\left(E^{\otimes 2}\right) .$$

And taking full expectation allows conclusion.

Partial Participation. This result is straightforward.

■

E.2 Variance and covariance of sketching

In this Subsection, we compute the expectation, the variance, and the covariance of sketching. In Appendix E.2.1, we give the proof principle of our computation, in Appendix E.2.2, we compute the expectation and the variance, and in Appendix E.2.3, we compute the covariance.

We thank Baptiste Goujaud (École polytechnique, CMAP) who greatly helped to prove the following.

E.2.1 Proof principle

Let y in $\mathbb{R}^d$ with $\|y\|^2 = 1$, and x in $\mathbb{R}^d$. By Definition 13, for Φ in $\mathbb{R}^{h\times d}$, we have $\mathcal{C}_\Phi(x) = \frac{1}{p}\Phi^\dagger\Phi x$ with $\Phi^\dagger = \Phi^\top(\Phi\Phi^T)^{-1}$ and $p = h/d$.
To compute the expectation, the variance, and the covariance of $\mathcal{C}_\Phi(x)$, the idea is to compute $\mathbb{E}[y^\top C_\Phi(x)]$ and $\mathbb{E}[(y^\top C_\Phi(x))^2]$ by establishing Equation (S32) which allows controlling the randomness of sketching by using Equation (S33). To establish Equation (S32), first observe that $pC_\Phi(\cdots)$ is a projector into a subspace of dimension h, indeed we have $(pC_\Phi \odot pC_\Phi)(x) = pC_\Phi(x)$. Then there exists a random matrix P in $\mathcal{O}_d$ s.t. $pC_\Phi(x) = P^\top J_h P x$. It leads to:

$$y^\top C_\Phi(x) = \frac{1}{p} y^\top P^\top J_h P x = \frac{1}{p}(Py)^\top J_h (Px)\,.$$

Now we note $X = Px/\|x\|$ and $Y = Py$, hence $y^\top C_\Phi(x) = \frac{\|x\|}{p} Y^\top J_h X$, and because P is in $\mathcal{O}_d$, we have:

$$\begin{cases} \|X\|^2 = 1 \\ \|Y\|^2 = \|y\|^2 = 1 \\ \langle X, Y\rangle = \langle x, y\rangle / \|x\|\,. \end{cases}$$

Furthermore, P is a random projector, it follows that X and Y are sampled uniformly from the zero-center sphere of radius 1; i.e. $X \sim \mathrm{Unif}(\mathcal{S}_d(0,1))$ and $Y \sim \mathrm{Unif}(\mathcal{S}_d(0,1))$. However, X and Y are not independent, this is why, we consider that $X \sim \mathrm{Unif}(\mathcal{S}_d(0,1))$ and write Y s.t. $Y = aX + bu$ with u a random vector in $\mathbb{R}^d$ of norm 1 orthogonal to X, that is to say, $u|X$ is uniformly sampled on a zero-centered hyper-sphere of radius 1 orthogonal to the vector X (see illustration on Figure S11). It comes that:

$$y^\top C_\Phi(x) = \frac{\|x\|}{p} Y^\top J_h X = \frac{\|x\|}{p}(aX^\top + bu^T) J_h X = \frac{\|x\|}{p}(aX^\top J_h X + bu^\top J_h X)\,. \tag{S32}$$

Observe that for any i, j in $\{1, \cdots, d\}$, X_i, X_j (resp. u_i, u_j) have the same law, it results to:

$$\forall (i,j) \in \{1,\cdots,d\}^2,\ \forall k \in \mathbb{N}, \qquad \mathbb{E}[X_i^k] = \mathbb{E}[X_j^k] \quad \text{and} \quad \mathbb{E}[u_i^k] = \mathbb{E}[u_j^k]\,. \tag{S33}$$

This property is the key to compute the expectation, the variance, and the covariance of sketching.
We now compute a and b. First, by definition, we have:

$$\frac{\langle x, y\rangle}{\|x\|} = \langle X, Y\rangle = a\,\|X\|^2 = a\,,$$

then we write that:

$$1 = \|Y\|^2 = \frac{\langle x, y\rangle^2}{\|x\|^4}\,\|X\|^2 + b^2\,\|u\|^2 = \frac{\langle x, y\rangle^2}{\|x\|^2} + b^2\,,$$

which gives $b = \sqrt{1 - \frac{\langle x,y\rangle^2}{\|x\|^2}}$.

At the end, we have: $Y = aX + bu = \frac{\langle x,y\rangle}{\|x\|} X + \sqrt{1 - \frac{\langle x,y\rangle^2}{\|x\|^2}}\, u$.

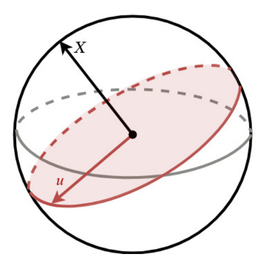


Figure S11: Sphere zero-center with radius 1: X and u are orthogonal.

E.2.2 Expectation and variance of sketching

In this Subsection, we prove that sketching verifies Item L.1 in Lemma 16; for this purpose, we show that it is unbiased, then we compute its variance.

Proposition S56 *Sketching is unbiased and its variance is relatively bounded, i.e., it verifies Item L.1 in Lemma 16 with $\omega = (1-p)/p$ where $p = h/d$.*

Proof Starting from Equation (S32), we have $y^\top C_\Phi(x) = \frac{\|x\|}{p}(aX^\top J_h X + bu^\top J_h X)$. We first compute the expectation w.r.t. the σ-algebra $\sigma(\{X\})$ generated by the noise involved in the random vector X, it gives:

$$\mathbb{E}[y^\top C_\Phi(x) \mid \sigma(\{X\})] = \frac{\|x\|}{p} \sum_{i=1}^{h} aX_i^2 + bX_i \mathbb{E}\left[u_i \mid \sigma(\{X\})\right] .$$

Because u is sampled uniformly from the zero-center sphere of radius 1 s.t. it is orthogonal to X, for any i in $\{1, \cdots, d\}$, we have $\mathbb{E}[u_i \mid \sigma(\{X\})] = 0$, hence taking full expectation, we obtain:

$$\mathbb{E}[y^\top C_\Phi(x)] = \frac{\|x\|}{p} \sum_{i=1}^{h} a\mathbb{E}[X_i^2] .$$

Using Equation (S33), we have $\mathbb{E}[X_i^2] = \frac{1}{d}\sum_{j=1}^d \mathbb{E}[X_j^2]$, next recalling that $p = h/d$ and $\|X\|^2 = 1$, it leads to $\mathbb{E}[y^\top C_\Phi(x)] = a\|x\|\mathbb{E}[\sum_{j=1}^d X_j^2] = a\|x\|\mathbb{E}[\|X\|^2] = a\|x\|$. And because $a = \langle x, y\rangle / \|x\|$, we have at the end that $\mathbb{E}[C_\Phi(x)] = x$. Now we compute the variance:

$$\mathbb{E}[C_\Phi(x)^\top C_\Phi(x)] = \frac{1}{p^2}\mathbb{E}[x^\top P^\top J_h P P^\top J_h P x] = \frac{1}{p^2}\mathbb{E}[x^\top P^\top J_h P x] = \frac{\|x\|^2}{p^2}\mathbb{E}[X^\top J_h X] .$$

$\mathbb{E}[X^\top J_h X]$ has been computed above and is equal to p, it results that $\mathbb{E}[C_\Phi(x)^\top C_\Phi(x)] = \|x\|^2 / p$. In the end, sketching verifies Lemma 16 with $\omega = (1-p)/p$. ■

E.2.3 Covariance of sketching.

In this Subsection, we compute the covariance of sketching. For the sake of demonstration, we need to compute the 4^{th}-moment of X_1 and the 2^{nd}-moment of u_1. For any i in $[d]$ and any vector v in $\mathbb{R}^d$, we note $v_{-i} = (v_j)_{j\in[d], j\neq i}$ in $\mathbb{R}^{d-1}$.

Computing the 4^{th}-moment of X_1.

The marginal density of X_1 is $f_{X_1} : x \mapsto B(\frac{d-1}{2}, \frac{1}{2})^{-1}(1-x^2)^{(d-3)/2}$ where B is the beta function defined as $B : x, y \mapsto \int_0^1 t^{x-1}(1-t)^{y-1} = 2\int_0^{\pi/2} \sin^{2x-1}(t)\cos^{2y-1}(t)\mathrm{d}t$. This result can be obtained either by an application of the formula for the surface area of a sphere (Li, 2010; Sidiropoulos, 2014), either by writing that $X_1 = \frac{Z_1}{\|Z\|}$ with Z a Gaussian vector with d components. Therefore we have that:

$$\mathbb{E}[X_1^4] = \frac{\int_{-1}^1 x^4(1-x^2)^{(d-3)/2}\mathrm{d}x}{2\int_0^{\pi/2}\sin^{d-2}(t)\mathrm{d}t} \overset{\text{(i)}}{=} \frac{2\int_0^{\pi/2}\cos^4(t)\sin^{d-2}(t)\mathrm{d}t}{2\int_0^{\pi/2}\sin^{d-2}(t)\mathrm{d}t} \overset{\text{(ii)}}{=} \frac{W_{d-2} - 2W_d + W_{d+2}}{W_{d-2}} ,$$

where at (i) we set $x = \cos(t)$ and at (ii) we make appears the Wallis' integrals defined for any n in $\mathbb{N}$ as $W_n = \int_0^{\pi/2} \sin^n(t) \mathrm{d}t$. Furthermore, we have the following recursion using integration by parts: $W_{d+2} = \frac{d+1}{d+2} W_d$, therefore, we have:

$$\mathbb{E}[X_1^4] = \left(1 - \frac{2(d-1)}{d} + \frac{(d-1)(d+1)}{d(d+2)}\right) = \frac{3}{d(d+2)}\,. \tag{S34}$$

Computing the 2^{nd}-moment of u_1 w.r.t the σ-algebra $\sigma(X)$.

We define three $(d-2)$–dimensional manifolds, two parallel hyperplanes P, P' and a sphere S, as follows:

$$\begin{cases} P = \{\tilde{u} \in \mathbb{R}^{d-1} \mid \langle \tilde{u}, X_{-i} \rangle = -X_i u_i\} \\ P' = \{\tilde{u} \in \mathbb{R}^{d-1} \mid \langle \tilde{u}, X_{-i} \rangle = 0\} \\ S = S_{d-1}(0, \sqrt{1-u_1^2}) \end{cases}$$

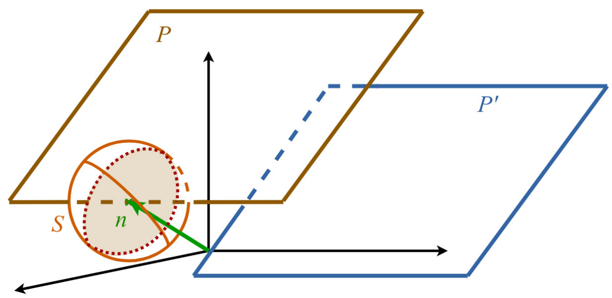


Figure S12: Parallel hyperplanes P and P' with the sphera S.

Obviously u_{-i} is in $P \cap S$; then we decompose u_{-i} in two terms $n + v$, with $v \sim \mathrm{Unif}(P')$ orthogonal to X and independent of u_i: n is the center of the sphere $S \cap P$ and v is its radius, n corresponds also to the normal vector of both P, P' with norm equal to the distance between the two hyperplanes, hence $n = \frac{\langle u_{-i}, X_{-i}\rangle}{\|X_{-i}\|^2} X_{-i} = -\frac{u_i X_i}{\|X_{-i}\|^2} X_{-i}$.

First, because $u_{-1} \in S$, we have $\|n + v\|^2 = 1 - u_1^2$, next by Pythagorean theorem this is equivalent to $\|v\|^2 = 1 - u_1^2 - \|n\|^2 = 1 - \frac{u_1^2}{\|X_{-1}\|^2}$. Second, because $u_{-1} \in P$, we have $u_1 = \frac{-\langle u_{-1}, X_{-1}\rangle}{X_1}$, that is to say the probability density function of $u_1 \mid X$ is proportional to the number of possible values for u_{-1}, which corresponds to the surface area of the hypersphere $P \cap S$. This surface is proportional to the radius $\|v\|^{d-4} = (1 - \frac{u_1^2}{\|X_{-1}\|^2})^{(d-4)/2}$ given that $P \cap S$ is a $(d-3)$–dimensional manifold, therefore:

$$\begin{aligned}\mathbb{E}[u_1^2 \mid \sigma(\{X\})] = \frac{\int_{-\|X_{-1}\|}^{\|X_{-1}\|} x^2 \left(1 - \frac{x^2}{\|X_{-i}\|^2}\right)^{(d-4)/2} \mathrm{d}x}{\int_{-\|X_{-1}\|}^{\|X_{-1}\|} \left(1 - \frac{x^2}{\|X_{-i}\|^2}\right)^{(d-4)/2} \mathrm{d}x} &\overset{\text{(i)}}{=} \frac{\|X_{-1}\|^2 \int_{-1}^{1} y^2 \left(1 - y^2\right)^{(d-4)/2} \mathrm{d}y}{\int_{-1}^{1} (1 - y^2)^{(d-4)/2} \mathrm{d}y} \\ &\overset{\text{(ii)}}{=} \|X_{-1}\|^2 \frac{W_{d-3} - W_{d-1}}{W_{d-3}}\,,\end{aligned}$$

where at (i) we set $y = \frac{x}{\|X_{-1}\|}$ and at (ii) we reuse the previous computations to make appear the Wallis' integral. In the end, we obtain:

$$\mathbb{E}[u_1^2 \mid \sigma(\{X\})] = (1 - \frac{d-2}{d-1}) \|X_{-1}\|^2 = \frac{\|X_{-1}\|^2}{d-1}\,. \tag{S35}$$

Note that this result is consistent with the fact that $\sum_{i=1}^d \mathbb{E}[u_i^2 \mid \sigma(\{X\})] = \frac{d - \sum_{i=1}^d X_i^2}{d-1} = 1$. Now we can compute the covariance of the sketching operator.

Proposition S57 *Let x in p_M, the covariance of sketching is equal to:*

$$\mathbb{E}[\mathcal{C}_\Phi(x)^{\otimes 2}] = \frac{1}{p}((\alpha - \beta)M + \beta \mathrm{Tr}\,(M)\,\mathrm{I}_d)\,,$$

with $\alpha = \frac{h+2}{d+2}$ and $\beta = \frac{d-h}{(d-1)(d+2)}$.

Proof
Let x in $\mathbb{R}^d$ and y in $\mathbb{R}^d$ with $\|y\|^2 = 1$, starting from Equation (S32), we have:

$$\begin{aligned}(y^\top C_\Phi(x))^2 &= \frac{\|x\|^2}{p^2}(aX^\top J_h X + bu^\top J_h X)^2 \\ &= \frac{\|x\|^2}{p^2}\left(a^2(X^\top J_h X)^2 + 2ab(X^\top J_h X u^\top J_h X) + b^2(u^\top J_h X)^2\right).\end{aligned}$$

First term. Taking expectation, we have $\mathbb{E}[(X^\top J_h X)^2] = \sum_{i=1}^h \left(\mathbb{E}[X_i^4] + \sum_{j=1, j\neq i}^h \mathbb{E}[X_i^2 X_j^2]\right)$. However:

$$\begin{aligned}\sum_{j=1,j\neq i}^h \mathbb{E}[X_i^2 X_j^2] = \mathbb{E}\left[X_i^2 \sum_{j=1,j\neq i}^h X_j^2\right] &\overset{\text{(i)}}{=} \mathbb{E}\left[X_i^2 \sum_{j=1,j\neq i}^h \frac{1}{d-1}\sum_{k=1,k\neq i}^d X_k^2\right] \\ &\overset{\text{(ii)}}{=} \frac{h-1}{d-1}\mathbb{E}\left[X_i^2(1-X_i^2)\right]\,,\end{aligned}$$

where we use at line (i) Equation (S33) and at line (ii) $\sum_{i=1}^d X_i^2 = 1$. It follows that:

$$\begin{aligned}\mathbb{E}[(X^\top J_h X)^2] &= \sum_{i=1}^h \left(\frac{d-h}{d-1}\mathbb{E}[X_i^4] + \frac{h-1}{d-1}\mathbb{E}\left[X_i^2\right]\right) \\ &\overset{\text{(i)}}{=} \frac{h(d-h)}{d-1}\mathbb{E}[X_1^4] + \frac{h-1}{d-1}\sum_{i=1}^h \mathbb{E}\left[X_i^2\right] \\ &\overset{\text{(iii)}}{=} \frac{h(d-h)}{d-1}\mathbb{E}[X_1^4] + \frac{h(h-1)}{d(d-1)} \\ &\overset{\text{eq. S34}}{=} \frac{3h(d-h)}{d(d-1)(d+2)} + \frac{h(h-1)}{d(d-1)} = \frac{h(h+2)}{d(d+2)} := \alpha'\,.\end{aligned}$$

Where we considered at line (i) that for any i in $\{1, \cdots, h\}$, $\mathbb{E}[X_i^4] = \mathbb{E}[X_1^4]$, and at line (ii) that $\sum_{i=1}^h \mathbb{E}\left[X_i^2\right] = \frac{h}{d}\mathbb{E}[\|X\|^2] = h/d$.
Second term. We compute the expectation w.r.t. the σ-algebra $\sigma(\{X\})$ generated by the noise involved in the random vector X. It gives $\mathbb{E}\left[X^\top J_h X u^\top J_h X \;\middle|\; \sigma(\{X\})\right] = 0$, because $u|X$ is uniformly sampled on a zero-centered hyper-sphere, and thus for any i in $\{1, \cdots, d\}$, we have $\mathbb{E}[u_i \mid \sigma(\{X\})] = 0$.
Third term. We have $(u^\top J_h X)^2 = \sum_{i=1}^h u_i^2 X_i^2 + \sum_{j=1,j\neq i}^h u_i u_j X_i X_j$. On one side, we compute the expectation w.r.t. the σ-algebra $\sigma(\{X\})$ generated by the noise involved in the random vector X:

$$\sum_{i=1}^h \mathbb{E}\left[u_i^2 X_i^2 \;\middle|\; \sigma(\{X\})\right] = \sum_{i=1}^h X_i^2 \mathbb{E}\left[u_i^2 \;\middle|\; \sigma(\{X\})\right] \overset{\text{eq. S35}}{=} \frac{1}{d-1}\sum_{i=1}^h X_i^2 \|X_{-i}\|^2\,.$$

Taking full expectation, we have $\sum_{i=1}^h \mathbb{E}[u_i^2 X_i^2] = \frac{1}{d-1}\sum_{i=1}^h \mathbb{E}[X_i^2(1-X_i^2)] = \frac{h}{d-1}(\frac{1}{d}-\mathbb{E}[X_1^4])$, because for any i in $\{1,\dots,h\}$, $\mathbb{E}[X_i^4] = \mathbb{E}[X_1^4]$ and $\sum_{i=1}^h \mathbb{E}\left[X_i^2\right] = \frac{h}{d}\mathbb{E}[\|X\|^2] = h/d$.
Let i in $[d]$, on the other side, we compute the expectation w.r.t. the σ-algebra $\sigma(\{X, u_i\})$ generated by the noise involved in the random vector X and the random variable u_i, hence we requires to compute $\mathbb{E}\left[u_j \mid \sigma(\{X, u_i\})\right]$. To do so, as before, we decompose u_{-i} in two terms $n+v$ (see Figure S12), with $v \sim \mathrm{Unif}(P')$ orthogonal to X and independent of u_i, hence $\mathbb{E}\left[v \mid \sigma(\{X, u_i\})\right] = 0$. It gives that $\mathbb{E}\left[u_{-i} \mid \sigma(\{X, u_i\})\right] = -\frac{u_i X_i}{\|X_{-i}\|^2} X_{-i}$. Thereby, replacing for any coordinate $j \neq i$ in $[d]$ the value of u_{-i} and taking expectation w.r.t. the σ-algebra $\sigma(\{X\})$, we obtain:

$$\sum_{i=1}^h \sum_{j=1, j\neq i}^h X_i X_j \mathbb{E}\left[u_i u_j \mid \sigma(\{X\})\right] = -\sum_{i=1}^h \sum_{j=1, j\neq i}^h \frac{1}{\|X_{-i}\|^2} X_i^2 X_j^2 \mathbb{E}\left[u_i^2 \mid \sigma(\{X\})\right]$$
$$\overset{\text{eq. S35}}{=} -\frac{1}{d-1}\sum_{i=1}^h \sum_{j=1, j\neq i}^h X_i^2 X_j^2$$
$$= -\frac{1}{d-1}\sum_{i=1}^h \sum_{j=1, j\neq i}^h X_i^2 \frac{1-X_i^2}{d-1}\,.$$

Finally, we have: $\sum_{i=1}^h \sum_{j=1, j\neq i}^h \mathbb{E}[X_i X_j u_i u_j] = -\frac{h(h-1)}{d(d-1)^2}(1-\sum_{i=1}^d \mathbb{E}[X_i^4])$. Putting together the two terms, we have that:

$$\mathbb{E}\left[(u^\top J_h X)^2\right] = \frac{h}{d-1}(\frac{1}{d} - \mathbb{E}[X_i^4]) - \frac{h(h-1)}{d(d-1)^2}(1 - d\mathbb{E}[X_1^4]) \overset{\text{eq. S34}}{=} \frac{h(d-h)}{d(d-1)(d+2)} := \beta'\,.$$

In the end, we have $\mathbb{E}[(y^\top C_\Phi(x))^2] = \frac{\|x\|^2}{p^2}(a^2\alpha' + b^2\beta')$. And because $\|y\|^2 = 1$, $a = \langle x, y\rangle / \|x\|$ and $b = \sqrt{1 - \langle x, y\rangle^2 / \|x\|^2}$, replacing them by their values gives:

$$y^\top \mathbb{E}[C_\Phi(x))^{\otimes 2}] y = \frac{\|x\|^2}{p^2}\left(\alpha' \frac{\langle x, y\rangle^2}{\|x\|^2} + \beta'\left(y^\top y - \frac{\langle x, y\rangle^2}{\|x\|^2}\right)\right)\,,$$

hence $\mathbb{E}[C_\Phi(x))^{\otimes 2}] = \frac{1}{p^2}\left((\alpha' - \beta') x x^\top + \beta' \|x\|^2 \mathrm{I}_d\right)$. To conclude, we consider that x is a random variable sampled from a distribution p_M, then taking expectation on this random variable we have: $\mathbb{E} C_\Phi(x)^{\otimes 2} = \frac{1}{p}\left((\alpha - \beta) M + \beta \mathrm{Tr}\,(M)\,\mathrm{I}_d\right)$, with $\alpha = \frac{\alpha'}{p} = \frac{h+2}{d+2}$ and $\beta = \frac{\beta'}{p} = \frac{d-h}{(d-1)(d+2)}$. ■

E.3 Proof of Propositions 24 and 25

In this Subsection, we give the proof of Propositions 24 and 25 which provides generic comparisons between the asymptotic convergence rate of compressors. We first give a lemma resulting from the Cauchy-Schwarz's inequality necessary to establish these proofs.

Lemma S58 (Cauchy-Schwarz's inequality on matrices' traces) *For any matrix M in $\mathbb{R}^{d\times d}$, we have $\mathrm{Tr}\,(M)\,\mathrm{Tr}\left(M^{-1}\right) \geq d^2$, with strict inequalities if M is not proportional to I_d. And if M is with constant diagonal equal to c in $\mathbb{R}$, we have $c\mathrm{Tr}\left(M^{-1}\right) \geq d$.*

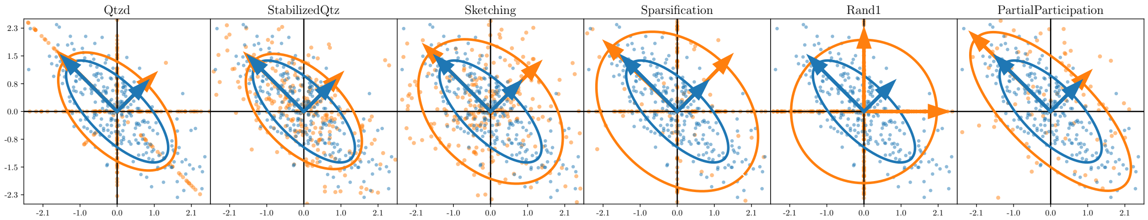


Figure S13: H not diagonal, scenario using features standardization. Scatter plot of $(x_k)_{i=1}^K$/ $(\mathcal{C}(x_k))_{i=1}^K$ with its ellipse $\mathcal{E}_{\mathrm{Cov}[x_k]}/\mathcal{E}_{\mathrm{Cov}[\mathcal{C}(x_k)]}$.

Proof Let M in $\mathbb{R}^{d\times d}$, using the Cauchy-Schwarz inequality, we have:

$$d^2 = \mathrm{Tr}\left(\mathrm{I}_d\right)^2 = \mathrm{Tr}\left(M^{1/2}M^{-1/2}\right)^2 \overset{\text{C.S}}{\leq} \mathrm{Tr}\left(M\right)\mathrm{Tr}\left(M^{-1}\right) ,$$

and we have equality if M is proportional to I_d. ■

Now we give the demonstration of Propositions 24 and 25. On Figure S13, we complete the numerical illustration provided in Subsection 3.3.1 by illustrating the scenario of standardized features, i.e., when the diagonal of M is the identity.

Proposition S59 (Comparison between $\mathcal{C}_{\mathrm{PP}}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_\Phi$, $\omega = d/h-1$) *We consider $\mathcal{C}$ in $\{\mathcal{C}_{\mathrm{PP}}, \mathcal{C}_{\mathrm{s}}, \mathcal{C}_{\mathrm{rd}h}, \mathcal{C}_\Phi\}$ with $p = h/d$, such that $\mathcal{C}$ always satisfies Lemma 16 with $\omega = d/h-1$. For any matrix $M \in \mathbb{R}^{d\times d}$:*

1. *If M is diagonal, then:*
   - $\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) = \mathfrak{C}(\mathcal{C}_{\mathrm{s}}, p_M) = \mathfrak{C}(\mathcal{C}_{\mathrm{rd}h}, p_M) = \frac{d}{h}M$,
   - $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\mathrm{PP/s/rd}h}, p_M)M^{-1}\right) \leq \mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_\Phi, p_M)M^{-1}\right)$.
2. *Moreover, for any matrix M with a* constant diagonal *(e.g., after standardization), we have:*

   $$\mathrm{Tr}(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M)M^{-1}) \leq \mathrm{Tr}(\mathfrak{C}(\mathcal{C}_\Phi, p_M)M^{-1}) \leq \mathrm{Tr}(\mathfrak{C}(\mathcal{C}_s, p_M)M^{-1}) \leq \mathrm{Tr}(\mathfrak{C}(\mathcal{C}_{\mathrm{rd}h}, p_M)M^{-1}) ,$$

   with strict inequalities if M is not proportional to I_d.

Proof
Let M in $\mathbb{R}^{d\times d}$ and take $p = h/d$.

Proof of Item 1 in Proposition 24. In the diagonal case, the first equalities are straightforward as we have $\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) = \mathfrak{C}(\mathcal{C}_{\mathrm{s}}, p_M) = \mathfrak{C}(\mathcal{C}_{\mathrm{rd}1}, p_M) = \frac{d}{h}M$. Next, we have (regardless if M is diagonal or not):

$$\begin{aligned}\mathrm{Tr}\left(\left(\mathfrak{C}(\mathcal{C}_\Phi, p_M) - \mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M)\right)M^{-1}\right) &= (\frac{h+1}{d+2} + \delta_{hd} - 1)\frac{\mathrm{Tr}\left(\mathrm{I}_d\right)}{p} + (1 - \frac{h-1}{d-1})\frac{\mathrm{Tr}\left(M\right)\mathrm{Tr}\left(M^{-1}\right)}{p(d+2)} \\ &\overset{\text{Lemma S58}}{\geq} \frac{d}{p}\left(\frac{h+1}{d+2} + \delta_{hd} - 1 + \frac{d}{d+2}(1 - \frac{h-1}{d-1})\right) \\ &= 0\,.\end{aligned}$$

Proof of Item 2 in Proposition 24. Suppose now that $\mathrm{Diag}\left(M\right) = c\mathrm{I}_d$, then we have $\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) = \frac{d}{h} M$, $\mathfrak{C}(\mathcal{C}_{\mathrm{s}}, p_M) = M + (\frac{d}{h} - 1)c\mathrm{I}_d$, $\mathfrak{C}(\mathcal{C}_{\mathrm{rd}h}, p_M) = \frac{d(h-1)}{h(d-1)} M + \frac{d}{h}(1 - \frac{h-1}{d-1})c\mathrm{I}_d$ and $\mathfrak{C}(\mathcal{C}_{\Phi}, p_M) = \frac{d}{h}\left((\frac{h+1}{d+2} - \delta_{hd})M + \left(1 - \frac{h-1}{d-1}\right) \frac{\mathrm{Tr}(M)}{d+2}\mathrm{I}_d \right)$. Firstly, from previous item, we have

$$\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) M^{-1}\right) \leq \mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\Phi}, p_M)\right) M^{-1} \,.$$

Secondly, we write:

$$\begin{aligned}
\mathrm{Tr}\left(\left(\mathfrak{C}(\mathcal{C}_{\Phi}, p_M) - \mathfrak{C}(\mathcal{C}_{\mathrm{s}}, p_M)\right) M^{-1}\right) &= \frac{d}{p}\left(\frac{h+1}{d+2} + \delta_{hd} - \frac{h}{d}\right) \\
&\quad + \frac{c\mathrm{Tr}\left(M^{-1}\right)}{p}\left(\frac{d}{d+2}(1 - \frac{h-1}{d-1}) - (1 - \frac{h}{d})\right) \\
&= \frac{d}{p}\left(\frac{h+1}{d+2} + \delta_{hd} - \frac{h}{d}\right) - \frac{c\mathrm{Tr}\left(M^{-1}\right)}{p} \cdot \frac{(d-2)(d-h)}{d(d-1)(d+2)} \\
&\overset{\text{Lemma S58}}{\leq} \frac{d}{p}\left(\frac{h+1}{d+2} + \delta_{hd} - \frac{h}{d} - \frac{(d-2)(d-h)}{d(d-1)(d+2)}\right) = 0 \,.
\end{aligned}$$

Thirdly, we have:

$$\begin{aligned}
\mathrm{Tr}\left(\left(\mathfrak{C}(\mathcal{C}_{\mathrm{rd}h}, p_M) - \mathfrak{C}(\mathcal{C}_{\mathrm{s}}, p_M)\right) M^{-1}\right) &= \frac{h-d}{h(d-1)}\mathrm{Tr}\left(\mathrm{I}_d\right) + \frac{d-h}{h(d-1)} c\mathrm{Tr}\left(M^{-1}\right) \\
&\overset{\text{Lemma S58}}{\geq} \frac{d}{h}\left(\frac{h-d}{d-1} + \frac{d-h}{d-1}\right) = 0 \,.
\end{aligned}$$

■

Proposition S60 (Comparison between $\mathcal{C}_{\mathrm{PP}}, \mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{s}}$, $\omega = \sqrt{d}$) *We consider $\mathcal{C}$ in $\{\mathcal{C}_{\mathrm{PP}}, \mathcal{C}_{\mathrm{q}}, \mathcal{C}_{\mathrm{s}}\}$ with $p = (\sqrt{d} + 1)^{-1}$, such that $\mathcal{C}$ always satisfies Lemma 16 with $\omega = \sqrt{d}$.*

1. *For any symmetric matrix M diagonal, we have:*

$$\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) M^{-1}\right) = \mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{s}, p_M) M^{-1}\right) \overset{possib.\ \ll}{\leq} \left(1 + \frac{1}{\sqrt{d}}\right) \mathrm{Tr}\left(\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) M^{-1}\right) \,.$$

2. *If M is not necessarily diagonal but with a* constant diagonal *(e.g., after standardization), then*

   - $\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) \preccurlyeq \mathfrak{C}(\mathcal{C}_{\mathrm{s}}, p_M)$
   - $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) M^{-1}\right) \leq \left(1 + \frac{1}{\sqrt{d}}\right) \mathrm{Tr}\left(\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) M^{-1}\right)$.

Proof
Let M in $\mathbb{R}^{d \times d}$ and take $p = \frac{1}{1+\sqrt{d}}$.

Proof of Item 1 in Proposition 25. In the diagonal case with $p = \frac{1}{1+\sqrt{d}}$, we have $\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) = \sqrt{\mathrm{Tr}\,(M)}\sqrt{M}$ and $\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) = (1+\sqrt{d})M$, hence $\mathrm{Tr}\left(\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M)M^{-1}\right) = \sqrt{\mathrm{Tr}\,(M)}\mathrm{Tr}\left(\sqrt{M^{-1}}\right)$ and $\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M)M^{-1}\right) = (1+\sqrt{d})d$. Noting $(\lambda_i)_{i\in[d]}$ the eigenvalues of M, and using the Cauchy-Schwarz inequality's, we have:

$$\begin{aligned}
d^2 = \left(\sum_{i=1}^d 1\right)^2 = \left(\sum_{i=1}^d \lambda_i^{1/4}\lambda_i^{-1/4}\right)^2 &\overset{\text{C.S}}{\leq} \left(\sum_{i=1}^d \lambda_i^{1/2}\right)\left(\sum_{i=1}^d \lambda_i^{-1/2}\right) \\
&\overset{\text{C.S}}{\leq} \sqrt{\sum_{i=1}^d \lambda_i}\sqrt{\sum_{i=1}^d 1}\left(\sum_{i=1}^d \lambda_i^{-1/2}\right) = \sqrt{d\mathrm{Tr}\,(M)}\mathrm{Tr}\left(M^{-1/2}\right) = \sqrt{d}\mathrm{Tr}\left(\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M)M^{-1}\right).
\end{aligned}$$

Which follows that $\mathrm{Tr}\left(\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M)M^{-1}\right) \geq d^{3/2} = \sqrt{d}(1+\sqrt{d})^{-1}\mathrm{Tr}\left(\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M)M^{-1}\right)$ and it allows to conclude.

Proof of Item 2 in Proposition 25. Suppose now that $\mathrm{Diag}\,(M) = c\mathrm{I}_d$, then we have $\mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) = (\sqrt{d}+1)M$, $\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) = M + (\sqrt{d}-1)c\mathrm{I}_d$, and $\mathfrak{C}(\mathcal{C}_{\mathrm{s}}, p_M) = M + c\sqrt{d}\mathrm{I}_d$. Firstly, it follows that:

$$\mathfrak{C}(\mathcal{C}_{\mathrm{s}}, p_M) - \widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) = \left(M + \sqrt{d}c\mathrm{I}_d\right) - \left(M + (\sqrt{d}-1)c\mathrm{I}_d\right) = c\mathrm{I}_d \succcurlyeq 0\,,$$

Secondly, we have $(1+\frac{1}{\sqrt{d}})\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) - \mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M) = -(1-\frac{1}{\sqrt{d}})M + (\sqrt{d}-\frac{1}{\sqrt{d}})c\mathrm{I}_d$, which gives:

$$\begin{aligned}
\mathrm{Tr}\left(\left((\sqrt{d}-\frac{1}{\sqrt{d}})\widetilde{\mathfrak{C}}(\mathcal{C}_{\mathrm{q}}, M) - \mathfrak{C}(\mathcal{C}_{\mathrm{PP}}, p_M)\right)M^{-1}\right) &= (\sqrt{d}-\frac{1}{\sqrt{d}})c\mathrm{Tr}\left(M^{-1}\right) - (1-\frac{1}{\sqrt{d}})\mathrm{Tr}\,(\mathrm{I}_d) \\
&\geq (\sqrt{d}-\frac{1}{\sqrt{d}})d - (1-\frac{1}{\sqrt{d}})d \quad \text{(Lemma S58)} \\
&\geq d(\sqrt{d}-1) \geq 0\,.
\end{aligned}$$

And the proof is concluded.

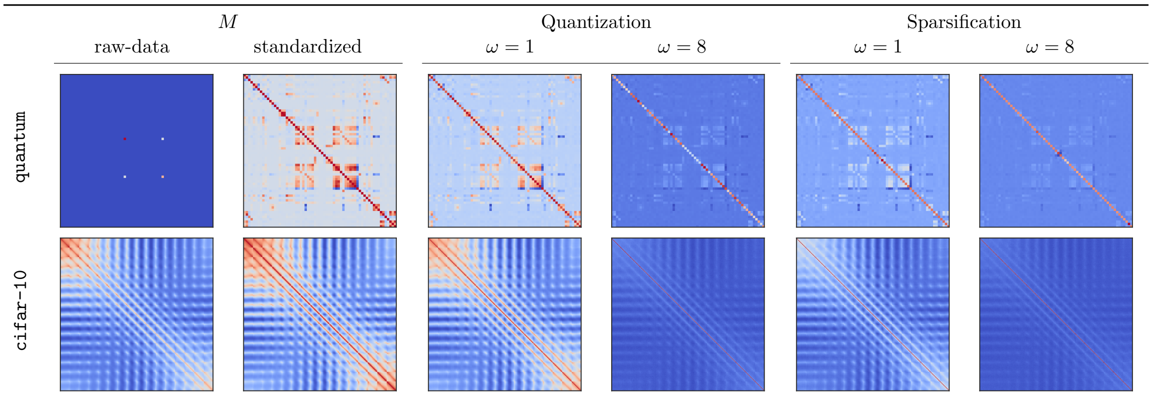
■

E.4 Empirical covariances computed on quantum and cifar10

On Table S3, for both `quantum` and `cifar-10`, we first plot the covariance matrix (1) without any processing and (2) with standardization. In this latter case, we then plot the covariances induced by quantization and sparsification for $\omega = 1$ and 8. For `quantum`, without standardization, only four points are visible; it is caused by some rows having extremely large values at features 27 and 43, resulting in a feature mean 100 times greater than the others.

Looking at the covariance induced by the compressors, we observe that for small ω, quantization better preserves the matrix structure compared to sparsification. This fact is consistent with Figure 6 where is given the trace of $\mathfrak{C}(\mathcal{C}_M, p_H)M^{-1}$ for these eight covariances: the traces for quantization are indeed smaller than for sparsification. This is also consistent with Figures 7c and 7f where $\omega = 1$ and where quantization outperforms sparsification.

Table S3: (1) Data covariances for `quantum` and `cifar-10`. (2) Covariance $\mathfrak{C}(\mathcal{C}_M, p_H)$ w./w.o. standardization for quantization and sparsification; see Figure 6 to have the corresponding trace of $\mathfrak{C}(\mathcal{C}_M, p_H)M^{-1}$.



Appendix F. Technical results on federated learning.

F.1 Validity of the assumptions made on the random fields in the case of covariate-shift

In this Subsection, we examine the setting of federated and compressed LSR under the scenario of covariate-shift (Subsection 4.1). Specifically, we consider the case where for any i, j in $[\![1, N]\!]$, we have heterogeneous covariances, i.e., $H_i \neq H_j$, but a unique optimal model i.e. $w_*^i = w_*$. We verify that all the assumptions on the random fields done in Subsection 2.1 are fulfilled in the setting. For this purpose, we redefine the filtration given in Appendix D to align them with the FL setting. For k in $\mathbb{N}^*$ and for i in $[N]$, we note u_k^i the noise that controls the compression $\mathcal{C}_k^i(\cdot)$ at round k.

Definition S61 *We note $(\mathcal{G}_k)_{k\in\mathbb{N}}$ the filtration associated with the features noise, $(\mathcal{H}_k)_{k\in\mathbb{N}}$ the filtration associated with the label noise, and $(\mathcal{I}_k)_{k\in\mathbb{N}}$ the filtration associated to the stochastic gradient noise, which is the union of the two previous filtrations. For $k \in \mathbb{N}^*$, we define $\mathcal{F}_0 = \{\varnothing\}$ and*

$$\begin{aligned}
\mathcal{G}_k &= \sigma\left(\mathcal{F}_{k-1} \cup \{(x_k^i)_{i=1}^N\}\right)\\
\mathcal{H}_k &= \sigma\left(\mathcal{F}_{k-1} \cup \{(\varepsilon_k^i)_{i=1}^N\}\right)\\
\mathcal{I}_k &= \sigma\left(\mathcal{F}_{k-1} \cup \{(x_k^i, \varepsilon_k^i)_{i=1}^N\}\right)\\
\mathcal{F}_k &= \sigma\left(\mathcal{F}_{k-1} \cup \{(x_k, \varepsilon_k^i, u_k^i)_{i=1}^N\}\right) .
\end{aligned}$$

Now we prove that all assumptions done in Section 2 are correct in this setting.

Property S62 (Validity of the setting presented in Definition 2) *For Algorithm 3 in the context of Model 1, we have that the setting presented in Definition 2 is verified.*

Proof From Algorithm 3, we have for any k in $\mathbb{N}^*$ and any w in $\mathbb{R}^d$, $\xi_k(w - w_*) = \nabla F(w) - \frac{1}{N}\sum_{i=1}^N \mathcal{C}_k^i(g_k^i(w))$. Because $(g_k^i)_{k\in\mathbb{N}^*, i\in[\![1,N]\!]}$ and $(\mathcal{C}_k^i)_{k\in\mathbb{N}^*, i\in[\![1,N]\!]}$ are by definition two sequences of i.i.d. random fields (Algorithm 3), it follows that their composition is also i.i.d., *hence $(\xi_k)_{k\in\mathbb{N}^*}$ is a sequence of i.i.d. random fields.*

Taking expectation w.r.t. the σ-algebra $\mathcal{I}_k$ we have $\mathbb{E}\left[\mathcal{C}_k^i(g_k^i(w)) \mid \mathcal{I}_k\right] = g_k^i(w)$ (Lemma 16), next with the σ-algebra $\mathcal{F}_{k-1}$, we have $\mathbb{E}\left[g_k^i(w) \mid \mathcal{F}_{k-1}\right] = \nabla F_i(w)$ (Equation (2)). And because $\frac{1}{N}\sum_{i=1}^N \nabla F_i(w) = \nabla F(w)$, *we obtain that the random fields are zero-centered.* From Model 1, we have for any k in $\mathbb{N}^*$ and any w in $\mathbb{R}^d$ that:

$$\begin{aligned}
F(w) &= \frac{1}{2N}\sum_{i=1}^N \mathbb{E}\left[(\langle x_k^i, w\rangle - y_k^i)^2\right] \\
&= \frac{1}{2N}\sum_{i=1}^N \mathbb{E}\left[(w-w_*)^\top (x_k^i \otimes x_k^i)(w-w_*) - 2\varepsilon_k^i \left\langle x_k^i, w-w_*\right\rangle + (\varepsilon_k^i)^2\right] \\
&= \frac{1}{2N}\sum_{i=1}^N (w-w_*)^\top H_i (w-w_*) + \sigma^2 = \frac{1}{2}((w-w_*)^\top \overline{H}(w-w_*) + \sigma^2)\,.
\end{aligned}$$

And we have from Model 1: $\mathrm{Tr}\left(\overline{H}\right) = \frac{1}{N}\sum_{i=1}^N \mathrm{Tr}\left(H_i\right) = \frac{1}{N}\sum_{i=1}^N R_i^2 =: \overline{R}^2$, which concludes the verification. ■

Property S63 (Validity of Assumption 1) *Consider Algorithm 3 and Model 1 with Lemma 16, for any iteration k in $\mathbb{N}^*$, the second moment of the additive noise ξ_k^{add} can be bounded by $(\omega+1)\overline{R}^2\sigma^2/N$ i.e. Assumption 1 is verified.*

Proof Let k in $\mathbb{N}^*$. Because we consider Algorithm 3, with Definitions 2 and 4, we first have $\xi_k^{\mathrm{add}} = -\frac{1}{N}\sum_{i=1}^N \mathcal{C}_k^i(g_{k,*}^i)$, hence taking expectation w.r.t the σ-algebra $\mathcal{I}_k$ and because the N compressions are independent (Algorithm 3), using Lemma 16, we have that:

$$\begin{aligned}
\mathbb{E}\left[\|\xi_k^{\mathrm{add}}\|^2 \;\middle|\; \mathcal{I}_k\right] &= \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left\|\mathcal{C}_k^i(g_{k,*}^i)\right\|^2 \;\middle|\; \mathcal{I}_k\right] + \frac{1}{N^2}\sum_{i\neq j}\left\langle g_{k,*}^i, g_{k,*}^j\right\rangle \\
&\leq \frac{\omega+1}{N^2}\sum_{i=1}^N \left\|g_{k,*}^i\right\|^2 + \frac{1}{N^2}\sum_{i\neq j}\left\langle g_{k,*}^i, g_{k,*}^j\right\rangle\,.
\end{aligned}$$

Next, we first have from Model 1 and Equation (2) that for any i in $[N]$, $g_{k,*}^i = -\varepsilon_k^i x_k^i$, secondly because $\left((\varepsilon_k^i)_{k\in[K],i\in[N]}\right)$ are independent from $\left((x_k^i)_{k\in[K],i\in[N]}\right)$ (Model 1), we have that $\mathbb{E}[\|\varepsilon_k^i x_k^i\|^2] \leq \sigma^2 R_i^2$, hence $\mathbb{E}\left[\|\xi_k^{\mathrm{add}}\|^2 \mid \mathcal{F}_{k-1}\right] = \mathbb{E}[\|\xi_k^{\mathrm{add}}\|^2] = \frac{\omega+1}{N^2}\sum_{i=1}^N \sigma^2 R_i^2\,.$ ■

Property S64 (Validity of Assumption 2.1) *Consider Algorithm 3 in the context of Model 1 with Lemma 16, for any iteration k in $\mathbb{N}^*$, the second moment of the multiplicative noise $\xi_k^{\mathrm{mult}}(w)$ can be bounded for any w in $\mathbb{R}^d$ by $2(\omega+1)\max_{i\in[N]}(R_i^2)\left\|\overline{H}^{1/2}(w-w_*)\right\|^2/N + 4(\omega+1)\overline{R}^2\sigma^2/N$ i.e. Assumption 2.1 is verified.*

Proof Let k in $\mathbb{N}^*$, we note $\eta = w - w_*$. Because we consider Algorithm 3, with Definitions 2 and 4, we write $\xi_k^{\mathrm{mult}}(\eta) = \frac{1}{N}\sum_{i=1}^N \xi_k^{i,\mathrm{mult}}(\eta)$, where $\xi_k^{i,\mathrm{mult}}(\eta) = H_i\eta - \mathcal{C}(g_k^i(w)) + \mathcal{C}(g_{k,*}^i)$ is the multiplicative noise on client i in $[N]$, hence developing the squared norm gives:

$$\left\|\xi_k^{\mathrm{mult}}(\eta)\right\|^2 = \left\|\frac{1}{N}\sum_{i=1}^N \xi_k^{i,\mathrm{mult}}(\eta)\right\|^2 = \frac{1}{N^2}\sum_{i=1}^N \left\|\xi_k^{i,\mathrm{mult}}(\eta)\right\|^2 + \frac{1}{N^2}\sum_{i\neq j}\left\langle \xi_k^{i,\mathrm{mult}}(\eta), \xi_k^{j,\mathrm{mult}}(\eta)\right\rangle .$$

Taking expectation w.r.t. the σ-algebra $\mathcal{F}_{k-1}$, using that the N compressions are independent (Algorithm 3) and that for any i in $[N]$, $\mathbb{E}[\xi_k^{i,\mathrm{mult}}(\eta) \mid \mathcal{F}_{k-1}] = 0$ (Lemma 16) results to have:

$$\mathbb{E}[\|\xi_k^{\mathrm{mult}}(\eta)\|^2 \mid \mathcal{F}_{k-1}] = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}[\|\xi_k^{i,\mathrm{mult}}(\eta)\|^2 \mid \mathcal{F}_{k-1}] .$$

Next, we use the result of Property S50 for each client i in $[N]$ and we obtain:

$$\begin{aligned}\mathbb{E}\left[\|\xi_k^{\mathrm{mult}}(\eta)\|^2 \;\middle|\; \mathcal{F}_{k-1}\right] &\leq \frac{1}{N^2}\sum_{i=1}^N \left(2(\omega+1)R_i^2\|H_i^{1/2}(w-w_*)\|^2 + 4(\omega+1)R_i^2\sigma^2\right)\\ &\leq \frac{2(\omega+1)\max_{i\in[N]}(R_i^2)}{N}\|\overline{H}^{1/2}(w-w_*)\|^2 + \frac{4(\omega+1)\overline{R}^2\sigma^2}{N} ,\end{aligned}$$

which allows concluding. ■

Property S65 (Validity of Assumption 2.2) *Consider Algorithm 3 in the context of Model 1 with Lemma 16, for any iteration k in $\mathbb{N}^*$, the second moment of the multiplicative noise $\xi_k^{\mathrm{mult}}(w)$ can be bounded for any w in $\mathbb{R}^d$ by $(\Omega\sigma\max_{i\in[N]}(R_i^2)\|\overline{H}^{1/2}(w-w_*)\| + (\omega+1)\max_{i\in[N]}(R_i^2)\|\overline{H}^{1/2}(w-w_*)\|^2)/N$ i.e. Assumption 2.2 is verified.*

Proof Let k in $\mathbb{N}^*$, we note $\eta = w - w_*$. From Property S64, taking expectation w.r.t. the σ-algebra $\mathcal{F}_{k-1}$, decomposing the multiplicative noise results to have:

$$\mathbb{E}[\|\xi_k^{\mathrm{mult}}(\eta)\|^2 \mid \mathcal{F}_{k-1}] = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}[\|\xi_k^{i,\mathrm{mult}}(\eta)\|^2 \mid \mathcal{F}_{k-1}] .$$

Next we use the result of Property S51 for each client i in $[N]$ and we obtain:

$$\mathbb{E}[\|\xi_k^{\mathrm{mult}}(\eta)\|^2 \mid \mathcal{F}_{k-1}] \leq \frac{1}{N^2}\sum_{i=1}^N \Omega R_i^2\sigma\sqrt{\|H_i^{1/2}(w-w_*)\|^2} + (\omega+1)R_i^2\|H_i^{1/2}(w-w_*)\|^2 .$$

With Jensen's inequality S7 used for concave function:

$$\begin{aligned}\mathbb{E}\left[\left\|\xi_k^{\mathrm{mult}}(\eta)\right\|^2 \;\middle|\; \mathcal{F}_{k-1}\right] &\leq \frac{\Omega\sigma\max_{i\in[N]}(R_i^2)}{N}\sqrt{\frac{1}{N}\sum_{i=1}^N\|H_i^{1/2}(w-w_*)\|^2} \\ &\quad+\frac{(\omega+1)\max_{i\in[N]}(R_i^2)}{N^2}\sum_{i=1}^N\|H_i^{1/2}(w-w_*)\|^2 \\ &\leq \frac{\Omega\sigma\max_{i\in[N]}(R_i^2)}{N}\sqrt{\|\overline{H}^{1/2}(w-w_*)\|^2} \\ &\quad+\frac{1}{N}(\omega+1)\max_{i\in[N]}(R_i^2)\|\overline{H}^{1/2}(w-w_*)\|^2\,,\end{aligned}$$

which allows concluding.

■

Property S66 (Validity of Assumption 3) *Consider Algorithm 3 and Model 1 with Lemma 16, if the compressor $\mathcal{C}$ is linear, then for any iteration k in $\mathbb{N}^*$, the multiplicative noise ξ_k^{mult} is linear, thus there exist a matrix Ξ_k in $\mathbb{R}^{d\times d}$ such that for any w in $\mathbb{R}^d$, $\xi_k^{\mathrm{mult}}(w)=\Xi_k w$. Furthermore the second moment of the multiplicative noise can be bounded for any w in $\mathbb{R}^d$ by $(\omega+1)\max_{i\in[N]}(R_i^2)\left\|\overline{H}^{1/2}(w-w_*)\right\|^2/N$, hence Assumption 3 is verified.*

Proof Let k in $\mathbb{N}^*$, we note $\eta=w-w_*$. Because we consider Algorithm 3, with Definitions 2 and 4, we write $\xi_k^{\mathrm{mult}}(\eta)=\frac{1}{N}\sum_{i=1}^N\xi_k^{i,\mathrm{mult}}(\eta)$, where $\xi_k^{i,\mathrm{mult}}(\eta)=H_i\eta-\mathcal{C}(g_k^i(w))+\mathcal{C}(g_{k,*}^i)$ is the multiplicative noise on client i in $[N]$. And because for any clients i in $\{1,\cdots N\}$ the random mechanism $\mathcal{C}_k^i$ is linear, there exists a random matrix Π_k^i in $\mathbb{R}^{d\times d}$ s.t. for any z in $\mathbb{R}^d$, we have $\mathcal{C}_k^i(z)=\Pi_k^i z$, it follows that:

$$\xi_k^{\mathrm{mult}}(\eta)=\nabla F(w)-\frac{1}{N}\sum_{i=1}^N\mathcal{C}_k^i(g_k^i(w))+\mathcal{C}_k^i(g_{k,*}^i)=\left(\overline{H}-\frac{1}{N}\sum_{i=1}^N\Pi_k^i(x_k^i\otimes x_k^i)\right)\eta\,.$$

Hence, the first part of Assumption 2.2 is verified with $\Xi_k=\frac{1}{N}\sum_{i=1}^N H_i-\Pi_k^i(x_k^i\otimes x_k^i)$. From Property S64, taking expectation w.r.t. the σ-algebra $\mathcal{F}_{k-1}$, decomposing the multiplicative noise results to have:

$$\mathbb{E}\left[\left\|\xi_k^{\mathrm{mult}}(\eta)\right\|^2 \;\middle|\; \mathcal{F}_{k-1}\right]=\frac{1}{N^2}\sum_{i=1}^N\mathbb{E}\left[\left\|\xi_k^{i,\mathrm{mult}}(\eta)\right\|^2 \;\middle|\; \mathcal{F}_{k-1}\right].$$

Next we use the result of Property S52 for each client i in $[N]$ and we obtain:

$$\begin{aligned}\mathbb{E}\left[\left\|\xi_k^{\mathrm{mult}}(\eta)\right\|^2 \;\middle|\; \mathcal{F}_{k-1}\right] &\leq \frac{1}{N}\sum_{i=1}^N(\omega+1)R_i^2\left\|H_i^{1/2}(w-w_*)\right\|^2 \\ &\leq \frac{(\omega+1)\max_{i\in[N]}(R_i^2)}{N^2}\left\|\frac{1}{N}\sum_{i=1}^N H_i^{1/2}(w-w_*)\right\|^2,\end{aligned}$$

which allows concluding.

■

Property S67 (Validity of Assumption 4) *Considering Algorithm 3 under the setting of Model 2 with Remark 1 and Lemma 16, if the compressor $\mathcal{C}$ is linear, then for any k in $\mathbb{N}^*$, we have $\mathfrak{C}_{\mathrm{ania}} \preccurlyeq \sigma^2 \max_{i\in[N]}(\text{Ш}_{H_i})\overline{H}/N$ and $\mathbb{E}\left[\Xi_k \Xi_k^\top\right] \preccurlyeq \max_{i\in[N]}(R_i^2 \text{Ш}_{H_i})\overline{H}/N$, with $(\text{Ш}_{H_i})_{i\in[N]}$ given in Corollary 18. Overall, Assumption 4 is thus verified.*

Proof
First inequality.
By Definition 6, we have $\mathfrak{C}_{\mathrm{ania}} = \mathbb{E}[\xi_k^{\mathrm{add}} \otimes \xi_k^{\mathrm{add}} \mid \mathcal{F}_{k-1}] = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}[\mathcal{C}_k^i(g_{k,*}^i)^{\otimes 2} \mid \mathcal{F}_{k-1}]$, because for any client i in $[N]$ $\left((\varepsilon_k^i)_{k\in[K]}\right)$ is independent from $\left((x_k^i)_{k\in[K]}\right)$ (Model 1) and using compressor linearity and Equation (S30), it gives:

$$\begin{aligned}\mathfrak{C}_{\mathrm{ania}} &= \sigma^2 \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\mathcal{C}_k^i(x_k^i)^{\otimes 2}\right] = \frac{\sigma^2}{N^2}\sum_{i=1}^N \mathfrak{C}(\mathcal{C}^i, p_{H_i}) \preccurlyeq \frac{\sigma^2}{N^2}\sum_{i=1}^N \text{Ш}_{H_i} H \\ &\preccurlyeq \frac{\sigma^2 \max_{i\in[N]}(\text{Ш}_{H_i})}{N}\overline{H}\,.\end{aligned}$$

Second inequality.
Using Property S66, because the random mechanism $\mathcal{C}^i$ is linear, there exists two matrices Π_k^i, Ξ_k^i in $\mathbb{R}^{d\times d}$ s.t. for any z in $\mathbb{R}^d$, we have $\mathcal{C}_k^i(z) = \Pi_k^i z$ and $\xi_k^{\mathrm{mult,i}}(z) = \Xi_k^i z = (H_i - \Pi_k^i(x_k^i \otimes x_k^i)z$, which gives that $\Xi_k = \frac{1}{N}\sum_{i=1}^N H_i - \Pi_k^i(x_k^i \otimes x_k^i)$. It follows that:

$$\Xi_k \Xi_k^\top = \frac{1}{N^2}\sum_{i=1}^N (\Xi_k^i)(\Xi_k^i)^\top + \frac{1}{N^2}\sum_{i\neq j}(\Xi_k^i)(\Xi_k^j)^\top\,.$$

Taking the σ-algebra $\mathcal{F}_{k-1}$, using that the N compressions are independent (Algorithm 3) and that for any i in $[N]$, $\mathbb{E}\left[\xi_k^{i,\mathrm{mult}} \,\middle|\, \mathcal{F}_{k-1}\right] = 0$ (Lemma 16) results to have $\mathbb{E}[\Xi_k \Xi_k^\top \mid \mathcal{F}_{k-1}] = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[(\Xi_k^i)(\Xi_k^i)^\top \,\middle|\, \mathcal{F}_{k-1}\right]$. Now, we can reuse the computations given in Property S53 to obtain $\mathbb{E}\left[(\Xi_k^i)(\Xi_k^i)^\top \,\middle|\, \mathcal{F}_{k-1}\right] \preccurlyeq R_i^2 \text{Ш}_{H_i} H_i$. Therefore, we have $\mathbb{E}\left[\Xi_k \Xi_k^\top \,\middle|\, \mathcal{F}_{k-1}\right] \preccurlyeq \max_{i\in[N]}(R_i^2 \text{Ш}_{H_i})\overline{H}/N$, which concludes the second part of the verification of Assumption 4.

■

F.2 Heterogeneous optimal point

In this section, we explore further the scenario of concept-shift by adding a memory mechanism (Mishchenko et al., 2019). This mechanism has been shown by Philippenko and Dieuleveut (2020) to improve the convergence in the case of heterogeneous clients. We give below the updates equation defining the algorithm of distributed compressed LSR with memory.

Algorithm 4 (Distributed compressed LMS with control variates) *Each client $i \in [N]$ maintains a sequence $(h_k^i)_{i\in[N]}$ in $\mathbb{R}^d$, observes at any step $k \in [K]$ an oracle $\mathrm{g}_k^i(\cdot)$ on the gradient of the local objective function F_i and applies an independent random compression mechanism $\mathcal{C}_k^i(\cdot)$ to the difference $\mathrm{g}_k^i - h_k^i$. And for any step-size $\gamma > 0$, any $k \in \mathbb{N}^*$, the sequence of iterates $(w_k)_{k\in\mathbb{N}}$ satisfies:*

$$\begin{cases} w_k = w_{k-1} - \dfrac{\gamma}{N} \displaystyle\sum_{i=1}^N \mathcal{C}_k^i(\mathrm{g}_k^i(w_{k-1}) - h_{k-1}^i) + h_{k-1}^i \\ h_k^i = h_{k-1}^i + \alpha \mathcal{C}_k^i(\mathrm{g}_k^i(w_{k-1}) - h_{k-1}^i)\,, \end{cases} \tag{S36}$$

with $\alpha = 1/2(\omega+1)$.

The counterpart of adding memory is that the random fields are no more identically distributed, thus Definition 2 is not fulfilled, and results from Section 2 cannot be applied, especially because $\mathbb{E}\left[\xi_k^{\mathrm{add}} \otimes \xi_k^{\mathrm{add}}\right]$ changes along iterations. To remedy this problem, we define here the *limit* of the covariance of the additive noise i.e. $\mathfrak{C}_{\mathrm{ania}}^\infty = \lim\limits_{k\to+\infty} \mathbb{E}\left[\xi_k^{\mathrm{add}} \otimes \xi_k^{\mathrm{add}}\right]$. In the following result, we establish an asymptotic result on the convergence, similar to Theorem 8.

Theorem S68 (CLT for concept-shift heterogeneity) *Consider Algorithm 4 under Model 1 with $\mu > 0$ and Lemma 16, 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 $g_{k,*}^i - \nabla F_i(w_*)$.*

Theorem S68 shows that when using memory, in the settings of heterogeneous optimal points $(w_*^i)_{i=1}^N$, convergence is still impacted by heterogeneity but with a smaller additive noise's covariance as $\Theta_i' \prec \Theta_i$. In particular, in the case of deterministic gradients (batch case), we case $\Theta_i' \equiv 0$. From a technical standpoint, it shows that we recover asymptotically the results stated by Theorems 8 and 12 in the general setting of i.i.d. random fields $(\xi_k(\eta_{k-1}))_{k\in\mathbb{N}^*}$. To prove this theorem, we show that the assumptions required by Theorem S36 are fulfilled by this framework.

Proof

For the sake of demonstration, we define a Lyapunov function V_k (as in Mishchenko et al., 2019; Liu et al., 2020; Philippenko and Dieuleveut, 2020), with k in $[\![1, K]\!]$:

$$V_k = \|\eta_k\|^2 + 2\gamma_k^2 C \frac{1}{N}\sum_{i=1}^N \left\|h_{k-1}^i - \nabla F_i(w_*)\right\|^2\,,$$

with C in $\mathbb{R}^*$ being a Lyapunov constant that verifies some conditions given in Theorem S6 in Philippenko and Dieuleveut (2020). For any k in $\mathbb{N}$, the Lyapunov function is defined combining two terms: (1) the distance from parameter w_k to the optimal parameter w_*, (2) for any client i in $[N]$, the distance between the memory term h_{k-1}^i and the true gradient $\nabla F_i(w_*)$.

First, we have that in the case of decreasing step size s.t. for any k in $\mathbb{N}$, $\gamma_k = k^{-\alpha}$, we have $\eta_K \xrightarrow[K\to+\infty]{L^2} 0$ and $h^i_K \xrightarrow[K\to+\infty]{L^2} \nabla F_i(w_*)$.
Let k in $\mathbb{N}^*$, from the demonstration of the `Artemis` algorithm with memory, we have from Theorem S6 in Philippenko and Dieuleveut (2020) (see page 41-45) that (1) combining Equations (S14) and (S15) in the form (S14)+$2\gamma_k^2 C$(S15), (2) and applying strong-convexity, allows to obtain Equation (S17) but adapted to decreasing step-size:

$$\mathbb{E}\left[V_k \mid \mathcal{F}_{k-1}\right] \leq (1 - 2\gamma_k\mu\square_k)\left\|w_{k-1} - w_*\right\|^2 + \frac{2\gamma_k^2 C\diamondsuit}{N}\sum_{i=1}^N \left\|h^i_{k-1} - \nabla F_i(w_*)\right\|^2 + \frac{2\gamma_k^2\sigma\triangle}{N}\,,$$

with $\square_k, \diamondsuit, \triangle$ being three constants in $\mathbb{R}$ whose exact expression can be found on pages 42-43 in Philippenko and Dieuleveut (2020). Furthermore, in the same article, they verify that to obtain a $(1-\gamma_k\mu)$ convergence, the following condition on $\square_k, \diamondsuit, \triangle$ are fulfilled for any k in $\mathbb{N}^*$: $\square_k \leq 1/2$ and $\diamondsuit \leq 1 - \gamma_k\mu$.
These properties are valid under some conditions on the Lyapunov constant C, the step-size γ_k, and the learning rate α; these conditions are provided in the statement of Theorem S6 from (Philippenko and Dieuleveut, 2020) and we don't recall them here. Hence, we can write that we have:

$$\mathbb{E}\left[V_k \mid \mathcal{F}_{k-1}\right] \leq (1 - \gamma_k\mu)\left(\left\|w_{k-1} - w_*\right\|^2 + \frac{2\gamma_k^2 C}{N}\sum_{i=1}^N \left\|h^i_{k-1} - \nabla F_i(w_*)\right\|^2\right) + \frac{2\gamma_k^2\sigma^2\triangle}{N}\,,$$

and because for any k in N, the step-size is decreasing, we have $\gamma_k \leq \gamma_{k-1}$, which makes to recover the Lyapunov function V_{k-1} at step $k-1$: $\mathbb{E}\left[V_k \mid \mathcal{F}_{k-1}\right] \leq (1-\gamma_k\mu)V_{k-1} + \frac{2\gamma_k^2\sigma^2\triangle}{N}$.
Taking full expectation and unrolling the sequence $(V_k)_{k\in\mathbb{N}}$, we obtain:

$$\mathbb{E}V_k \leq \prod_{i=1}^k (1-\gamma_i\mu)V_0 + \frac{2\sigma^2\triangle}{N}\sum_{j=1}^k \gamma_j^2 \prod_{i=j+1}^k (1-\gamma_i\mu)\,.$$

To show that each part of the bound given in the above equation tends to zero when k grows to infinity is classical computations and can be find for instance in lectures notes of Bach (2022, pages 164-167 and 182), and Kushner and Yin (2003).
To apply Theorem 1 from Polyak and Juditsky (1992, recalled in Theorem S36), which gives the desired result, it suffices to prove the convergence in probability of the covariance of the noise $\xi_k(\eta_{k-1})$ towards $\mathfrak{C}_{\mathrm{ania}}$, as $k\to\infty$.
In the following, we show that $\lim_{k\to+\infty} \mathbb{E}\left[\xi_k(\eta_{k-1})\xi_k(\eta_{k-1})^\top \;\middle|\; \mathcal{F}_{k-1}\right] \overset{\mathbb{P}}{=} \mathfrak{C}^\infty_{\mathrm{ania}}$**.** Let k in $\mathbb{N}^*$, for this purpose, we consider the following additive/multiplicative noise decomposition:

$$\begin{cases} \xi^{\mathrm{A}}_{k,*} = -\dfrac{1}{N}\displaystyle\sum_{i=1}^N \mathcal{C}^i_k(g^i_{k,*} - \nabla F_i(w_*)) \\ \xi^{\mathrm{M}}_k(\eta_k) = H_F\eta_k - \dfrac{1}{N}\displaystyle\sum_{i=1}^N \mathcal{C}^i_k(g^i_k(w_{k-1}) - h^i_{k-1}) + \dfrac{1}{N}\displaystyle\sum_{i=1}^N \mathcal{C}^i_k(g_{k,*} - \nabla F_i(w_*)) + h^i_{k-1}\,. \end{cases} \tag{S37}$$

Furthermore, we have that $\xi_k^{\mathrm{add}} \xrightarrow[k\to+\infty]{L^2} \xi_{k,*}^{\mathrm{A}}$ because of the Hölder-inequality (Lemma 16) and because we shown that $h_K^i \xrightarrow[K\to+\infty]{L^2} \nabla F_i(w_*)$; thereby $\mathbb{E}[\xi_k^{\mathrm{add}} \otimes \xi_k^{\mathrm{add}}] \xrightarrow[k\to+\infty]{L^1} \mathfrak{C}_{\mathrm{ania}}^{\infty}$. Next, from Equation (S37), we write:

$$
\begin{aligned}
\xi_k(\eta_{k-1})\xi_k(\eta_{k-1})^\top &= (\xi_{k,*}^{\mathrm{A}} - \xi_k^{\mathrm{M}}(\eta_{k-1}))(\xi_{k,*}^{\mathrm{A}} - \xi_k^{\mathrm{M}}(\eta_{k-1}))^\top \\
&= \xi_{k,*}^{\mathrm{A}} \otimes \xi_{k,*}^{\mathrm{A}} - \xi_{k,*}^{\mathrm{A}} \xi_k^{\mathrm{M}}(\eta_{k-1})^\top - \xi_k^{\mathrm{M}}(\eta_{k-1})(\xi_{k,*}^{\mathrm{A}})^\top + \xi_k^{\mathrm{M}}(\eta_{k-1}) \otimes \xi_k^{\mathrm{M}}(\eta_{k-1})\,.
\end{aligned}
$$

(i) Developing the covariance of the additive noise and considering Model 1 and Algorithm 3 results to $\mathbb{E}[\xi_{k,*}^{\mathrm{A}} \otimes \xi_{k,*}^{\mathrm{A}} \mid \mathcal{F}_{k-1}] = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}[\mathcal{C}_k^i(g_{k,*}^i - \nabla F_i(w_*))^{\otimes 2} \mid \mathcal{F}_{k-1}]$. For any iteration k in $\mathbb{N}^*$ and any client i in $[N]$, we note Θ_i' the covariance of $g_{k,*}^i - \nabla F_i(w_*)$, then $g_{k,*}^i - \nabla F_i(w_*)$ is an i.i.d. zero-centered random vectors draw from a distribution $p_{\Theta_i'}$, hence we have for any iteration k in $\mathbb{N}^*$, $\mathfrak{C}_{\mathrm{ania}}^{\infty} = \mathbb{E}[\xi_{k,*}^{\mathrm{A}} \otimes \xi_{k,*}^{\mathrm{A}} \mid \mathcal{F}_{k-1}] = \overline{\mathfrak{C}(\mathcal{C}^i, (p_{\Theta_i'})_{i=1}^N)}\,$.
(ii) Second, we show that $\mathbb{E}\left[\xi_k^{\mathrm{M}}(\eta_{k-1})^{\otimes 2} \;\middle|\; \mathcal{F}_{k-1}\right]$ converge to 0 in probability: it is sufficient to show that $\|\xi_k^{\mathrm{M}}(\eta_{k-1})^{\otimes 2}\|_F$ tends to 0. To do so, we use the fact that $\|\xi_k^{\mathrm{M}}(\eta_{k-1})^{\otimes 2}\|_F = \|\xi_k^{\mathrm{M}}(\eta_{k-1})\|_2^2$, it results to the following decomposition:

$$
\begin{aligned}
\|\xi_k^{\mathrm{M}}(\eta_{k-1})^{\otimes 2}\| \leq 3\,\|H\eta_{k-1}\|^2 &+ 3\left\|\frac{1}{N}\sum_{i=1}^N \mathcal{C}_k^i(g_k^i(w_{k-1}) - h_{k-1}^i) - \mathcal{C}_k^i(g_{k,*}^i - \nabla F_i(w_*))\right\|^2 \\
&+ 3\left\|\frac{1}{N}\sum_{i=1}^N h_{k-1}^i - \nabla F_i(w_*)\right\|^2\,.
\end{aligned}
$$

Considering the Hölder inequality given in Item L.2 from Lemma 16, because $\eta_k \xrightarrow[k\to+\infty]{L^2} 0$ and $h_k^i \xrightarrow[k\to+\infty]{L^2} \nabla F_i(w_*)$, we deduce that $\mathbb{E}\left[\xi_k^{\mathrm{M}}(\eta_{k-1})^{\otimes 2} \;\middle|\; \mathcal{F}_{k-1}\right]$ tends to 0 in L^1-norm.
(iii) Third, it remains to show that $\mathbb{E}[\xi_k^{\mathrm{M}}(\eta_{k-1})(\xi_{k,*}^{\mathrm{A}})^\top \mid \mathcal{F}_{k-1}] \xrightarrow[k\to+\infty]{\mathbb{P}} 0$. We use the Cauchy-Schwarz inequality's S8 for conditional expectation:

$$
\begin{aligned}
\mathbb{E}\left[\xi_k^{\mathrm{M}}(\eta_{k-1})(\xi_{k,*}^{\mathrm{A}})^\top\|_F \;\middle|\; \mathcal{F}_{k-1}\right]^2 &= \mathbb{E}\left[\xi_k^{\mathrm{M}}(\eta_{k-1})\|_2\|(\xi_{k,*}^{\mathrm{A}})^\top\|_2 \;\middle|\; \mathcal{F}_{k-1}\right]^2 \\
&\leq \|\mathbb{E}\left[\xi_k^{\mathrm{M}}(\eta_{k-1})\|_2^2 \;\middle|\; \mathcal{F}_{k-1}\right] \mathbb{E}\left[\|(\xi_{k,*}^{\mathrm{A}})^\top\|_2^2 \;\middle|\; \mathcal{F}_{k-1}\right]\,.
\end{aligned}
$$

The sequence of random vectors $(\xi_{k,*}^{\mathrm{A}})_{k\in\mathbb{N}}$ is i.i.d., and moreover we have shown previously that $\xi_k^{\mathrm{M}}(\eta_{k-1})^{\otimes 2}$ tends to 0, hence $\mathbb{E}[\xi_k^{\mathrm{M}}(\eta_{k-1})(\xi_{k,*}^{\mathrm{A}})^\top \mid \mathcal{F}_{k-1}]$ converges to 0 in distribution. Consequently, noting $\Theta_i' = \mathbb{E}[g_{k,*}^i - \nabla F_i(w_*))^{\otimes 2}]$ we can state that:

$$
\mathbb{E}\left[\xi_k(\eta_{k-1})^{\otimes 2} \;\middle|\; \mathcal{F}_{k-1}\right] \xrightarrow[k\to+\infty]{\mathbb{P}} \mathfrak{C}_{\mathrm{ania}}^{\infty} = \overline{\mathfrak{C}(\mathcal{C}^i, (p_{\Theta_i'})_{i=1}^N)}\,.
$$

■