ON THE SAGA ALGORITHM WITH DECREASING STEP*

BERNARD BERCU†, LUIS FREDES †, AND EMÉRIC GBAGUIDI †

Abstract. Stochastic optimization naturally appear in many application areas, including machine learning. Our goal is to go further in the analysis of the Stochastic Average Gradient Accelerated (SAGA) algorithm. To achieve this, we introduce a new λ-SAGA algorithm which interpolates between the Stochastic Gradient Descent ($\lambda = 0$) and the SAGA algorithm ($\lambda = 1$). Firstly, we investigate the almost sure convergence of this new algorithm with decreasing step which allows us to avoid the restrictive strong convexity and Lipschitz gradient hypotheses associated to the objective function. Secondly, we establish a central limit theorem for the λ-SAGA algorithm. Finally, we provide the non-asymptotic $\mathbf{L}^p$ rates of convergence.

Key words. SAGA algorithm, decreasing step, almost sure convergence, asymptotic normality, non-asymptotic rates of convergence

MSC codes. 68Q25, 68T05, 60F05, 60F15

1. Introduction.

Our goal is to solve the classical optimization problem in $\mathbb{R}^d$ which can be written as

$$(\mathcal{P}) \qquad \min_{x \in \mathbb{R}^d} f(x),$$

where f is the average of many functions,

$$f(x) = \frac{1}{N} \sum_{k=1}^{N} f_k(x). \tag{1.1}$$

This type of problem is frequently encountered in statistical learning and a standard way to solve $(\mathcal{P})$ is to make use of the Gradient Descent algorithm. However, in a large N context, this approach has a very high computational cost. This limitation has led to the development of many stochastic algorithms for optimization [8, 26].

These new methods have taken a major role in recent advances of the neural networks. Our goal is to go further in the analysis of the Stochastic Gradient Descent (SGD) algorithm [33] and the SAGA algorithm [12]. The standard SGD algorithm is given for all $n \geqslant 1$, by

$$\text{(SGD)} \qquad X_{n+1} = X_n - \gamma_n \nabla f_{U_{n+1}}(X_n) = X_n - \gamma_n (\nabla f(X_n) + \varepsilon_{n+1}),$$

where the initial state X_1 is a squared integrable random vector of $\mathbb{R}^d$ which can be arbitrarily chosen, $\nabla f(X_n)$ is the gradient of the function f calculated at the value X_n, $\varepsilon_{n+1} = \nabla f_{U_{n+1}}(X_n) - \nabla f(X_n)$ and (U_n) is a sequence of independent and identically distributed random variables, with uniform distribution on $\{1, 2, \ldots, N\}$, which is also independent from the sequence (X_n). Moreover, (γ_n) is a positive deterministic sequence decreasing towards zero and satisfying the standard conditions

$$\sum_{n=1}^{\infty} \gamma_n = +\infty \qquad \text{and} \qquad \sum_{n=1}^{\infty} \gamma_n^2 < +\infty. \tag{1.2}$$

*Submitted to the editors September 11, 2025.

Funding: This project has benefited from state support managed by the Agence Nationale de la Recherche (French National Research Agency) under the reference ANR-20-SFRI-0001.

†Institut de Mathématiques de Bordeaux, Université de Bordeaux (bernard.bercu@math.u-bordeaux.fr, luis.fredes@math.u-bordeaux.fr, thierry-emeric.gbaguidi@math.u-bordeaux.fr).

We clearly have from (1.1) that (ε_n) is a martingale difference sequence adapted to the filtration $(\mathcal{F}_n)$ where $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$.
The SAGA algorithm is a stochastic variance reduction algorithm which was proposed ten years ago in the pioneering work of [12]. It slightly differs from the SGD algorithm as it is given, for all $n \geqslant 1$, by

$$\text{(SAGA)} \qquad X_{n+1} = X_n - \gamma_n \left(\nabla f_{U_{n+1}}(X_n) - g_{n,U_{n+1}} + \frac{1}{N} \sum_{k=1}^{N} g_{n,k} \right),$$

where the initial states X_0 and X_1 are squared integrable random vectors of $\mathbb{R}^d$ which can be arbitrarily chosen, the initial value $g_{1,k}$ is given, for any $k = 1, \ldots, N$, by $g_{1,k} = \nabla f_k(X_0)$. Moreover, the sequence $(g_{n,k})$ is updated, for all $n \geqslant 1$ and $1 \leqslant k \leqslant N$, as

$$g_{n+1,k} = \begin{cases} \nabla f_k(X_n) & \text{if } U_{n+1} = k, \\ g_{n,k} & \text{otherwise.} \end{cases} \tag{1.3}$$

One can observe that in most of all papers dealing with the SAGA algorithm, the step size is a fixed value γ which depends on the strong convexity constant μ and the Lipschitz gradient constant L associated with f. This will not be the case here at all as the benefit of decreasing step algorithms has been recently investigated in [23]. Our work aims to investigate the almost sure convergence as well as the asymptotic normality of the SGD and SAGA algorithms with decreasing step sequence (γ_n) satisfying (1.2).

Our contributions. The goal of this paper is to answer to several natural questions.

(a) Is it possible to study the convergence of the SAGA algorithm with decreasing step ?
(b) Can we relax the strong convexity and the Lipschitz gradient assumptions ?
(c) Can we prove a central limit theorem for our new version of the SAGA algorithm ?
(d) Is it possible to provide non-asymptotic $\mathbf{L}^p$ bounds for the SAGA algorithm ?

We shall propose positive answers to all these questions by extending [12] in several directions.

The paper is structured as follows. Section 2 is devoted to the state of the art concerning the SGD and SAGA algorithms. In Section 3, we present our new version of the SAGA algorithm which we shall call the λ-SAGA algorithm. Section 4 deals with the main results of the paper. We establish the asymptotic properties of our λ-SAGA algorithm such as the almost sure convergence and the asymptotic normality. Non-asymptotic $\mathbf{L}^p$ rates of convergence are also provided. In Section 5, we illustrate our theoretical results by some numerical experiments on real dataset. All technical proofs are postponed to the appendices.

2. Related work. The stochastic approximations, initiated by [33] and [20], have taken a major role in optimization issues. The SGD algorithm, often known as a special case of the Robbins-Monro algorithm, is probably the most standard stochastic algorithm used in machine learning. The properties of this algorithm were investigated in several studies. The almost sure convergence results were established in [7, 8, 13, 22, 34, 35, 37]. The convergence rates were proven in [21, 25, 26, 28]. The

study of the asymptotic normality of stochastic approximations also appear in several works such that [13, 14, 29, 36, 40, 41].

In a high-dimensional context, many accelerated algorithms were proposed in literature in order to improve the Robbins-Monro algorithm performances [1, 30, 15, 12, 38, 24]. In this paper, we will focus on the SAGA algorithm first introduced by [12] for the minimization of the average of many functions and which is a well-known variance reduction method. This algorithm is a variant of the Stochastic Average Gradient (SAG) method proposed earlier in [35, 37]. It uses the concept of covariates to make an unbiased variant of the SAG method that has similar performances but is easier to implement [17]. The idea behind the SAGA algorithm, is to make use of the control variates, a well-known technique in Monte-Carlo simulation designed to reduce the variance of the SGD algorithm in order to accelerate its convergence. This algorithm incorporates knowledge about gradients on all previous data points rather than only using the gradient for the sampled data point [12, 27]. This method requires a storage linear in N [16]. Several works have studied the convergence of the SAGA algorithm, which is undoubtedly one of the most celebrated variance reduction algorithms.

Defazio et al. [12] established that the SAGA algorithm converges in $\mathbf{L}^2$ at exponential rate. This result has been shown by assuming that the function f is μ-strongly convex and with L-Lipschitz gradient and by considering a fixed constant step γ which tightly depends on the unknown values μ and L. The almost sure convergence of the SAGA algorithm was not investigated in [12]. More recently, it was shown by [31] that for a fixed constant step $\gamma = 1/(3L)$, $f(X_n)$ and X_n both converge almost surely to $f(x^*)$ and x^* respectively, where x^* is the unique point of $\mathbb{R}^d$ such that $\nabla f(x^*) = 0$. This algorithm has been also investigated in [11, 16, 27, 32] and there are now many variations on the original SAGA algorithm of [12]. For example, [32] proposed a variant of the SAGA algorithm that includes arbitrary importance sampling and minibatching schemes.

Despite a decade of research, several issues remain open on the SAGA algorithm. The choice of the step γ_n is clearly one of them. The vast majority of the theory for the SAGA algorithm relies on a fixed constant step γ depending on the values μ and L [12, 11, 27, 16, 31, 17]. However, from a practical point of view, the values μ and L are unknown and there is no guarantee on the convergence results established for this algorithm. We shall propose here to make use of decreasing step sequence (γ_n) which allows us to avoid these constraints and relax some classic assumptions such that the μ-strong convexity. Moreover, to the best of our knowledge, no result about the asymptotic normality of the SAGA algorithm is available in the literature so far.

3. The λ-SAGA algorithm. We introduce in this section the λ-SAGA algorithm which can be seen as a generalization of the SAGA algorithm. We recall below the general principle of the Monte Carlo method that gave birth to the λ-SAGA algorithm. Suppose that we would like to estimate the expectation $\mathbb{E}[X]$ of a square integrable real random variable X. Let us also consider another square integrable real random variable Y strongly positively correlated to X and for which we know how to compute the expectation $\mathbb{E}[Y]$. Then, it is possible to find a reduced variance estimator of $\mathbb{E}[X]$, given by $Z_\lambda = X - \lambda(Y - \mathbb{E}[Y])$ with λ in $[0, 1]$ (see, e.g., [9, 12]). One can obviously see that $\mathbb{E}[Z_\lambda] = \mathbb{E}[X]$, which means that Z_λ is an unbiased estimator of $\mathbb{E}[X]$. Moreover, $\mathbb{V}[Z_\lambda] = \mathbb{V}[X] + \lambda^2\mathbb{V}[Y] - 2\lambda\mathbb{C}ov(X, Y)$. Hence, as soon as $\mathbb{C}ov(X, Y) > 0$, we can choose λ in $[0, 1]$ such that $\mathbb{V}[Z_\lambda] \leqslant \mathbb{V}[X]$. Now, using this

principle of variance reduction, the λ-SAGA algorithm is defined, for all $n \geqslant 1$, by

$$X_{n+1} = X_n - \gamma_n \left(\nabla f_{U_{n+1}}(X_n) - \lambda \left(g_{n,U_{n+1}} - \frac{1}{N} \sum_{k=1}^{N} g_{n,k} \right) \right), \tag{λ-SAGA}$$

where the initial states X_0 and X_1 are squared integrable random vectors of $\mathbb{R}^d$ which can be arbitrarily chosen, the parameter λ belongs to $[0, 1]$, and (γ_n) is a positive deterministic sequence decreasing towards zero and satisfying (1.2).

One can establish a link between the SGD, SAGA and λ-SAGA algorithms. Indeed, the λ-SAGA algorithm with $\lambda = 0$ corresponds to the absence of variance reduction and reduces to the SGD algorithm. Furthermore, one can easily see that we find again the SAGA algorithm by choosing $\lambda = 1$. The motivation to introduce and study the λ-SAGA algorithm comes from our desire to propose a unified convergence analysis for the SGD and SAGA algorithms and to investigate what happens in the intermediate cases $0 < \lambda < 1$. We shall now state the general assumptions which we will use in all the sequel.

ASSUMPTION 3.1. *Assume that function f is continuously differentiable with a unique equilibrium point x^* in $\mathbb{R}^d$ such that $\nabla f(x^*) = 0$.*

ASSUMPTION 3.2. *Suppose that for all $x \in \mathbb{R}^d$ with $x \neq x^*$,*

$$\langle x - x^*, \nabla f(x) \rangle > 0.$$

ASSUMPTION 3.3. *Assume there exists a positive constant L such that, for all $x \in \mathbb{R}^d$,*

$$\frac{1}{N} \sum_{k=1}^{N} \|\nabla f_k(x) - \nabla f_k(x^*)\|^2 \leqslant L \|x - x^*\|^2.$$

These assumptions are not really restrictive and they are fulfilled in many applications. One can observe that Assumption 3.2 is obviously weaker than the standard hypothesis that each function f_k for $1 \leqslant k \leqslant N$ is μ-strongly convex with $\mu > 0$. Note also that Assumption 3.3 ensures that at x^*, the gradient of all functions f_k for any $1 \leqslant k \leqslant N$, does not change arbitrarily with respect to the vector $x \in \mathbb{R}^d$. Such an assumption is essential for convergence of most gradient-based algorithms; without it, the gradient would not provide a good indicator of how far to move to decrease f. One can also observe that if each function f_k has Lipschitz continuous gradient with constant $\sqrt{L_k}$, then Assumption 3.3 is satisfied by taking L as the average value of all L_k. The most interesting improvement here is that both conditions are local in x^* and sufficient for all of our analysis.

4. Main results. In this section, we present the main results of the paper. First of all, we provide an almost sure convergence analysis for the λ-SAGA algorithm with decreasing step. After that, we establish its asymptotic normality. Lastly, we conclude this section by focusing on non-asymptotic $\mathbf{L}^p$ rates of convergence of this stochastic algorithm.

4.1. Almost sure convergence. Our first result deals with the almost sure convergence of the λ-SAGA algorithm.

THEOREM 4.1. *Consider a fixed $\lambda \in [0,1]$. Assume that (X_n) is the sequence generated by the λ-SAGA algorithm with decreasing step sequence (γ_n) satisfying (1.2). In addition, suppose that Assumptions 3.1, 3.2 and 3.3 are satisfied. Then, we have*

$$\lim_{n\to+\infty} X_n = x^* \qquad a.s. \tag{4.1}$$

and

$$\lim_{n\to+\infty} f(X_n) = f(x^*) \qquad a.s. \tag{4.2}$$

Proof. Recall that for all $n \geqslant 1$,

$$X_{n+1} = X_n - \gamma_n \left(\nabla f_{U_{n+1}}(X_n) - \lambda \left(g_{n,U_{n+1}} - \frac{1}{N}\sum_{k=1}^{N} g_{n,k} \right) \right).$$

Hence, the λ-SAGA algorithm can be rewritten as

$$X_{n+1} = X_n - \gamma_n \left(Y_{n+1} - \lambda Z_{n+1} \right), \tag{4.3}$$

where

$$\left\{ \begin{array}{ll} Y_{n+1} & = \nabla f_{U_{n+1}}(X_n), \\ \\ Z_{n+1} & = \nabla f_{U_{n+1}}(\phi_{n,U_{n+1}}) - \dfrac{1}{N}\displaystyle\sum_{k=1}^{N} \nabla f_k(\phi_{n,k}), \end{array} \right.$$

and $\phi_{n,k}$ is the point such that $g_{n,k} = \nabla f_k(\phi_{n,k})$. As $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ and the sequence (U_n) is independent of the sequence (X_n), we clearly have from (1.1) that

$$\mathbb{E}[Y_{n+1}|\mathcal{F}_n] = \nabla f(X_n) \qquad \text{and} \qquad \mathbb{E}[Z_{n+1}|\mathcal{F}_n] = 0 \qquad a.s. \tag{4.4}$$

As a consequence, (Z_n) is a martingale difference sequence adapted to the filtration $(\mathcal{F}_n)$.
Hereafter, define for all $n \geqslant 1$,

$$V_n = \|X_n - x^*\|^2.$$

We obtain from (4.3) that for all $n \geqslant 1$,

$$\begin{aligned} V_{n+1} &= \|X_{n+1} - x^*\|^2, \\ &= \|X_n - x^* - \gamma_n(Y_{n+1} - \lambda Z_{n+1})\|^2, \\ &= V_n - 2\gamma_n \langle X_n - x^*, Y_{n+1} - \lambda Z_{n+1} \rangle + \gamma_n^2 \|Y_{n+1} - \lambda Z_{n+1}\|^2. \end{aligned}$$

Moreover, we have from Jensen's inequality and the fact that λ belongs to $[0,1]$ that

$$\begin{aligned} &\mathbb{E}[\|Y_{n+1} - \lambda Z_{n+1}\|^2 | \mathcal{F}_n] \\ &= \mathbb{E}[\|(Y_{n+1} - \nabla f_{U_{n+1}}(x^*)) - \lambda(Z_{n+1} - \nabla f_{U_{n+1}}(x^*)) + (1-\lambda)\nabla f_{U_{n+1}}(x^*)\|^2 | \mathcal{F}_n] \\ &\leqslant 2\mathbb{E}[\|Y_{n+1} - \nabla f_{U_{n+1}}(x^*)\|^2 | \mathcal{F}_n] + 2\mathbb{E}[\|Z_{n+1} - \nabla f_{U_{n+1}}(x^*)\|^2 | \mathcal{F}_n] \\ &+ 2\mathbb{E}[\|\nabla f_{U_{n+1}}(x^*)\|^2 | \mathcal{F}_n]. \end{aligned} \tag{4.5}$$

First of all, we clearly have

$$\mathbb{E}[\|\nabla f_{U_{n+1}}(x^*)\|^2|\mathcal{F}_n] = \frac{1}{N}\sum_{k=1}^{N}\|\nabla f_k(x^*)\|^2 \qquad a.s. \tag{4.6}$$

In addition, denote

$$A_n = \frac{1}{N}\sum_{k=1}^{N}\|\nabla f_k(\phi_{n,k}) - \nabla f_k(x^*)\|^2 \qquad \text{and} \qquad \Sigma_n = \frac{1}{N}\sum_{k=1}^{N}\nabla f_k(\phi_{n,k}).$$

Since $\nabla f(x^*) = 0$, we obtain by expanding the norm that

$$\mathbb{E}[\|Z_{n+1} - \nabla f_{U_{n+1}}(x^*)\|^2|\mathcal{F}_n] = A_n - \|\Sigma_n\|^2 \qquad a.s. \tag{4.7}$$

Furthermore, define for all $x \in \mathbb{R}^d$,

$$\tau^2(x) = \frac{1}{N}\sum_{k=1}^{N}\|\nabla f_k(x) - \nabla f_k(x^*)\|^2.$$

One can observe that

$$\mathbb{E}[\|Y_{n+1} - \nabla f_{U_{n+1}}(x^*)\|^2|\mathcal{F}_n] = \tau^2(X_n) \qquad a.s. \tag{4.8}$$

Putting together the three contributions (4.6), (4.7) and (4.8), we deduce from (4.5) that

$$\mathbb{E}[\|Y_{n+1} - \lambda Z_{n+1}\|^2|\mathcal{F}_n] \leqslant 2(\tau^2(X_n) + A_n + \theta^*) \qquad a.s. \tag{4.9}$$

where

$$\theta^* = \frac{1}{N}\sum_{k=1}^{N}\|\nabla f_k(x^*)\|^2.$$

Consequently, it follows from (4.4) and (4.9) that for all $n \geq 1$,

$$\mathbb{E}[V_{n+1}|\mathcal{F}_n] \leqslant V_n - 2\gamma_n\langle X_n - x^*, \nabla f(X_n)\rangle + 2\gamma_n^2(\tau^2(X_n) + A_n + \theta^*) \quad a.s. \tag{4.10}$$

Furthermore, let (T_n) be the sequence of Lyapunov functions defined, for all $n \geqslant 2$, by

$$T_n = V_n + 2N\gamma_{n-1}^2 A_n. \tag{4.11}$$

It follows from the very definition of the sequence $(\phi_{n,k})$ associated with (1.3) that

$$\begin{aligned}\mathbb{E}[A_{n+1}|\mathcal{F}_n] &= \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}[\|\nabla f_k(\phi_{n+1,k}) - \nabla f_k(x^*)\|^2|\mathcal{F}_n],\\ &= \frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{N}\|\nabla f_k(X_n) - \nabla f_k(x^*)\|^2 + \left(1 - \frac{1}{N}\right)\|\nabla f_k(\phi_{n,k}) - \nabla f_k(x^*)\|^2\right),\\ &= \frac{1}{N}\tau^2(X_n) + \left(1 - \frac{1}{N}\right)A_n, \end{aligned} \tag{4.12}$$

almost surely. Hence, we obtain from (4.10) and (4.12) that

$$\begin{aligned}
\mathbb{E}[T_{n+1}|\mathcal{F}_n] &= \mathbb{E}[V_{n+1}|\mathcal{F}_n] + 2N\gamma_n^2\mathbb{E}[A_{n+1}|\mathcal{F}_n], \\
&\leqslant V_n - 2\gamma_n\langle X_n - x^*, \nabla f(X_n)\rangle + 2\gamma_n^2(\tau^2(X_n) + A_n + \theta^*) + 2N\gamma_n^2\mathbb{E}[A_{n+1}|\mathcal{F}_n] \\
&= V_n + 2N\gamma_n^2 A_n - 2\gamma_n\langle X_n - x^*, \nabla f(X_n)\rangle + 2\gamma_n^2(2\tau^2(X_n) + \theta^*), \\
&\leqslant V_n + 2N\gamma_{n-1}^2 A_n - 2\gamma_n\langle X_n - x^*, \nabla f(X_n)\rangle + 2\gamma_n^2(2\tau^2(X_n) + \theta^*), \\
&\leqslant T_n - 2\gamma_n\langle X_n - x^*, \nabla f(X_n)\rangle + 2\gamma_n^2(2\tau^2(X_n) + \theta^*). \qquad (4.13)
\end{aligned}$$

Additionally, we clearly have $V_n \leqslant T_n$ almost surely and it follows from Assumption 3.3 that

$$\tau^2(X_n) \leqslant LV_n \leqslant LT_n.$$

Finally, we deduce from (4.13) that

$$\mathbb{E}[T_{n+1}|\mathcal{F}_n] \leqslant (1 + 4L\gamma_n^2)T_n - 2\gamma_n\langle X_n - x^*, \nabla f(X_n)\rangle + 2\gamma_n^2\theta^* \qquad a.s., \qquad (4.14)$$

which can be rewritten as

$$\mathbb{E}[T_{n+1}|\mathcal{F}_n] \leqslant (1 + a_n)T_n + \mathcal{A}_n - \mathcal{B}_n \qquad a.s.$$

where $a_n = 4L\gamma_n^2$, $\mathcal{A}_n = 2\gamma_n^2\theta^*$ and $\mathcal{B}_n = 2\gamma_n\langle X_n - x^*, \nabla f(X_n)\rangle$. The four sequences (T_n), (a_n), $(\mathcal{A}_n)$ and $(\mathcal{B}_n)$ are positive sequences of random variables adapted to $(\mathcal{F}_n)$. We clearly have from (1.2) that

$$\sum_{n=1}^{\infty} a_n < +\infty \qquad \text{and} \qquad \sum_{n=1}^{\infty} \mathcal{A}_n < +\infty.$$

Then, it follows from the Robbins-Siegmund Theorem [34] given by Theorem A.1 that (T_n) converges a.s. towards a finite random variable T and the series

$$\sum_{n=1}^{\infty} \mathcal{B}_n < +\infty \qquad a.s. \qquad (4.15)$$

Consequently, (V_n) also converges a.s. to a finite random variable V. It only remains to show that $V = 0$ almost surely. Assume by contradiction that $V > 0$. For some positive constants $a < b$, denote by Ω the annulus of $\mathbb{R}^d$,

$$\Omega = \{x \in \mathbb{R}^d, \quad 0 < a < \|x - x^*\|^2 < b\}.$$

Let F be the function defined, for all $x \in \mathbb{R}^d$, by

$$F(x) = \langle x - x^*, \nabla f(x)\rangle.$$

We have from Assumption 3.1 that F is a continuous function in Ω compact. It implies that there exists a positive constant c such that $F(x) > c$ for all $x \in \Omega$. However, for n large enough, $X_n \in \Omega$, which ensures that $\gamma_n\langle X_n - x^*, \nabla f(X_n)\rangle > c\gamma_n$. Consequently, it follows from (4.15) that

$$\sum_{n=1}^{\infty} \gamma_n < +\infty.$$

This is of course in contradiction with assumption (1.2). Finally, we obtain that $V = 0$ almost surely, leading to

$$\lim_{n\to+\infty} X_n = x^* \qquad a.s.$$

By continuity of the function f, we also have (4.2), which completes the proof of Theorem 4.1. ∎

4.2. Asymptotic normality. We now focus our attention on the asymptotic normality of the λ-SAGA algorithm with decreasing step. In this subsection, we assume that f is twice differentiable and we denote by $H = \nabla^2 f(x^*)$ the Hessian matrix of f at the point x^*.

ASSUMPTION 4.2. *Suppose that f is twice differentiable with a unique equilibrium point x^* in $\mathbb{R}^d$ such that $\nabla f(x^*) = 0$. Denote by $\rho = \lambda_{min}(H)$ the minimum eigenvalue of H. We assume that $\rho > 1/2$.*

The central limit theorem for the λ-SAGA algorithm is as follows.

THEOREM 4.3. *Consider a fixed $\lambda \in [0, 1]$. Let (X_n) be the sequence generated by the λ-SAGA algorithm with decreasing step $\gamma_n = 1/n$. Suppose that Assumption 4.2 is satisfied. Assume also that*

$$\lim_{n\to+\infty} X_n = x^* \qquad a.s. \tag{4.16}$$

Then, we have the asymptotic normality

$$\sqrt{n}(X_n - x^*) \quad \xrightarrow[n\to+\infty]{\mathcal{L}} \quad \mathcal{N}_d(0, \Sigma) \tag{4.17}$$

where the asymptotic covariance matrix is given by

$$\Sigma = (1-\lambda)^2 \int_0^\infty (e^{-(H-\mathbb{I}_d/2)u})^T \Gamma e^{-(H-\mathbb{I}_d/2)u} du,$$

with

$$\Gamma = \frac{1}{N}\sum_{k=1}^{N} \nabla f_k(x^*) \left(\nabla f_k(x^*)\right)^T.$$

Proof. The proof of Theorem 4.3 can be found in Appendix B. ∎

Remark 4.4. It was proven in Theorem 4.1 that the almost sure convergence (4.16) holds under Assumptions 3.1, 3.2 and 3.3. It is obvious to see that Assumption 4.2 implies Assumption 3.1. Consequently, Theorem 4.3 is also true when replacing (4.16) by Assumptions 3.2 and 3.3.

Many conclusions can be drawn from Theorem 4.3. First of all, if we assume that Γ is a positive definite matrix, we obtain that our λ-SAGA algorithm with $0 \leqslant \lambda < 1$, converges towards a centered normal distribution with positive definite variance. However, as soon as $\lambda = 1$, the limit distribution becomes a centered normal with variance $\Sigma = 0$, in other words a Dirac measure. Thus, the asymptotic distribution of the SAGA algorithm has zero variance and one can therefore try to understand it. In fact, the conditional variances of the two terms of the martingale difference (ε_n) extracted from this algorithm, converge almost surely to exactly the same matrix. Therefore, the conditional variance of (ε_n) vanishes which explains the final result for the SAGA algorithm. Moreover, Theorem 4.3 clearly shows the

asymptotic variance reduction effect. Indeed, when λ grows to 1, we observe that the variance Σ decreases and converges towards 0. Hence, for statistical inference purposes such that hypothesis test and confidence interval, we can take λ just a little smaller than 1 to reduce the variance with respect to SGD, but without canceling it.

4.3. Non-asymptotic convergence rates. In the same vein as [4] for the Robbins-Monro algorithm, we shall now establish non-asymptotic $\mathbf{L}^p$ convergence rates. Hence, our goal is to investigate, for all integer $p \geqslant 1$, the convergence rate of $\mathbb{E}[\|X_n - x^*\|^{2p}]$ for the λ-SAGA algorithm where the decreasing step is defined, for all $n \geqslant 1$ by,

$$\gamma_n = \frac{c}{n^\alpha}, \tag{4.18}$$

for some positive constant c and $1/2 < \alpha \leqslant 1$. First of all, we focus our attention on the standard case $p = 1$ by analyzing our algorithm with a little more stringent condition than Assumption 3.2.

ASSUMPTION 4.5. *Assume there exists a positive constant μ such that for all $x \in \mathbb{R}^d$ with $x \neq x^*$,*

$$\langle x - x^*, \nabla f(x) \rangle \geqslant \mu \|x - x^*\|^2.$$

Although this is a strengthened version of Assumption 3.2, it is still weaker than the usual $\mu-$strong convexity assumption on the function f. This condition is sometimes called in the literature the Restricted Secant Inequality [18, 19, 39].

THEOREM 4.6. *Consider a fixed $\lambda \in [0, 1]$. Let (X_n) be the sequence generated by the λ-SAGA algorithm with decreasing step sequence (γ_n) defined by (4.18). Suppose that Assumptions 3.1, 3.3 and 4.5 are satisfied with $2c\mu \leqslant 2^\alpha$ and $2c\mu > 1$ if $\alpha = 1$. Then, there exists a positive constant K such that for all $n \geqslant 1$,*

$$\mathbb{E}\big[\|X_n - x^*\|^2\big] \leqslant \frac{K}{n^\alpha}. \tag{4.19}$$

Proof. The proof of Theorem 4.6 can be found in Appendix C. ∎

Next, we carry out our analysis in the general case $p \geqslant 1$. It requires a strengthened version of Assumption 3.3 given as follows.

ASSUMPTION 4.7. *Assume that for some integer $p \geqslant 1$, there exists a positive constant L_p such that for all $x \in \mathbb{R}^d$,*

$$\frac{1}{N} \sum_{k=1}^{N} \|\nabla f_k(x) - \nabla f_k(x^*)\|^{2p} \leqslant L_p \|x - x^*\|^{2p}.$$

THEOREM 4.8. *Consider a fixed $\lambda \in [0, 1]$. Let (X_n) be the sequence generated by the λ-SAGA algorithm with decreasing step sequence (γ_n) defined by (4.18) and such that the initial state X_1 belongs to $\mathbf{L}^{2p}$. Suppose that Assumptions 3.1, 4.5 and 4.7 are satisfied with $pc\mu \leqslant 2^\alpha$ and $c\mu > 1$ if $\alpha = 1$. Then, there exists a positive constant K_p such that for all $n \geqslant 1$,*

$$\mathbb{E}\big[\|X_n - x^*\|^{2p}\big] \leqslant \frac{K_p}{n^{p\alpha}}. \tag{4.20}$$

Proof. The proof of Theorem 4.8 can be found in Appendix D. ∎

Remark 4.9. It is easy to see that Assumption 4.7 also implies that for all $x \in \mathbb{R}^d$,

$$(f(x) - f(x^*))^p \leqslant \frac{\sqrt{L_p}}{2^p} \|x - x^*\|^{2p}. \tag{4.21}$$

Then, it follows from Theorem 4.8 together with inequality (4.21) that there exists a positive constant $M_p = 2^{-p} K_p \sqrt{L_p}$ such that for all $n \geqslant 1$,

$$\mathbb{E}\big[(f(X_n) - f(x^*))^p\big] \leqslant \frac{M_p}{n^{p\alpha}}. \tag{4.22}$$

5. Numerical experiments. Consider the logistic regression model (see, e.g., [3, 6]) associated with the classical minimization problem $(\mathcal{P})$ of the convex function f given, for all $x \in \mathbb{R}^d$, by

$$f(x) = \frac{1}{N} \sum_{k=1}^{N} f_k(x) = \frac{1}{N} \sum_{k=1}^{N} \left(\log(1 + \exp(\langle x, w_k \rangle)) - y_k \langle x, w_k \rangle \right),$$

where $x \in \mathbb{R}^d$ is a vector of unknown parameters, $w_k \in \mathbb{R}^d$ is a vector of features and the binary output $y_k \in \{0, 1\}$. As stated, this problem is equivalent to the log-likelihood maximization problem, where the aim is to find the parameter x^* that maximizes the probability of a given sample $((w_1, y_1), \ldots, (w_N, y_N))$, which follows a model depending only on the unknown parameter x. To be more precise, our model has a Bernoulli probability with parameter $p_k(x)$ following a logistic function for each k,

$$p_k(x) = \frac{\exp(\langle x, w_k \rangle)}{1 + \exp(\langle x, w_k \rangle)}.$$

It is easy to see that f is twice differentiable and its Hessian matrix is given by

$$\nabla^2 f(x) = \frac{1}{N} \sum_{k=1}^{N} p_k(x)(1 - p_k(x)) w_k w_k^T.$$

Consequently, f has an unique equilibrium point x^* and if we assume that the minimum eigenvalue of $H = \nabla^2 f(x^*)$ is greater than $1/2$, Assumption 4.2 will be automatically satisfied, and therefore Assumption 3.1 too. Moreover, one can observe that Assumptions 3.3 and 4.7 hold with

$$L_p = \frac{1}{4^p N} \sum_{k=1}^{N} ||w_k||^{4p}.$$

We conducted experiments on the MNIST dataset in order to present a visualisation of the almost sure convergence in Theorem 4.1, the asymptotic normality in Theorem 4.3 and the $\mathbf{L}^2$ bound in Theorem 4.6. For the almost sure convergence, the training database considered here includes $N = 60000$ images in gray-scale format and size 28×28. Each image w_k is therefore a vector of dimension $d = 28 \times 28 = 784$. Each of these images is identified with a number from 0 to 9 and we divide it into a binary classification so that $y_k = 0$ if $\{0, 1, 2, 3, 4\}$ and $y_k = 1$ if $\{5, 6, 7, 8, 9\}$. The results

concerning the convergence of the estimator X_n of x^* are illustrated in Figure 1. The convergence is ordered from slowest to fastest in an increasing order with respect to $\lambda \in \{0, 0.5, 0.9, 1\}$.

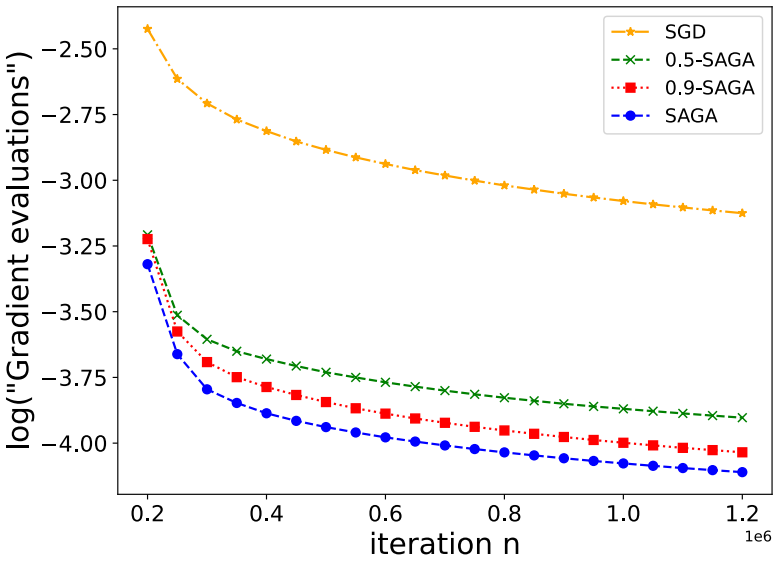


Fig. 1: Convergence with $\gamma_n = 1/n$ for 1.2M of iterations. Here we put "Gradient evaluations" since instead of using $\|\nabla f(X_n)\|$, we use the norm of the mean associated to the lines in the matrix g_n, $\|\sum_{k=1}^{N} g_{n,k}\|/N$. This quantity keeps track of the convergence since it also converges to 0 and its lines converge to the gradients of the functions f_k, that is for each $1 \leqslant k \leqslant N$, $\lim g_{n,k} = \nabla f_k(x^*)$ as n goes to infinity.

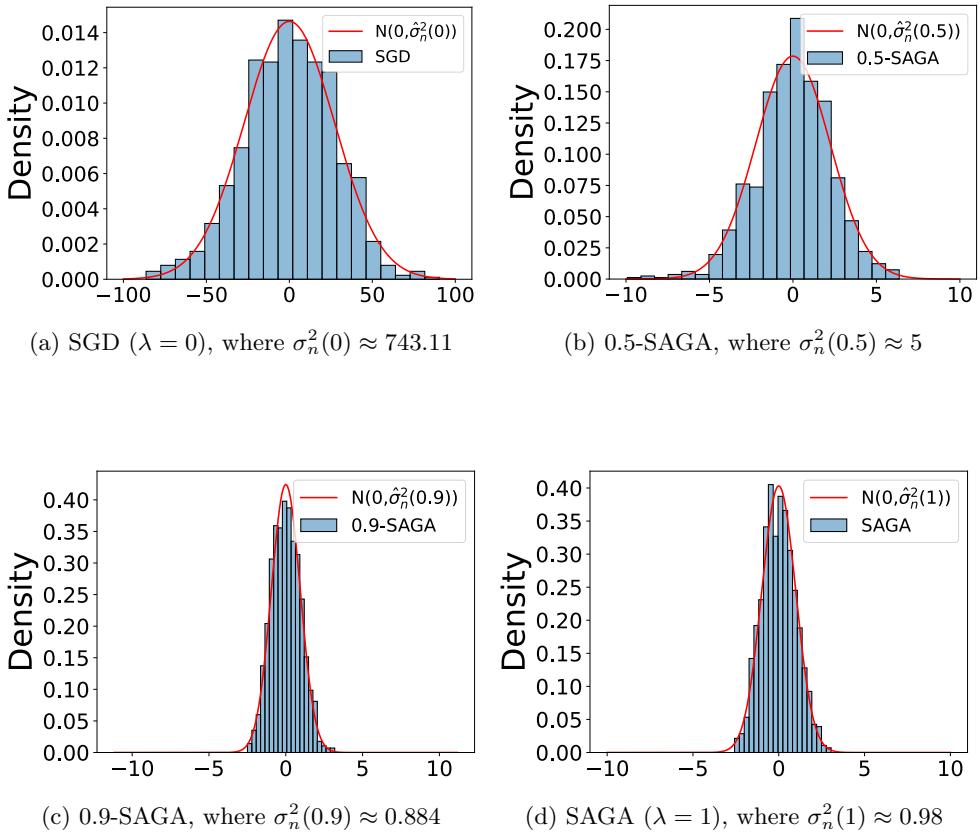
Moreover, to illustrate the asymptotic normality result, we use $N = 100$, the first 100 images in the MNIST dataset, and the distributional convergence

$$\lim_{n\to\infty} \mathbb{E}(h(\sqrt{n}(X_n - x^*))) = \mathbb{E}(h(\mathcal{N}_d(0, \Sigma))),$$

where h is defined, for all $x \in \mathbb{R}^d$, by $h(x) = \sum_{i=1}^{d} x_i$. It follows from Theorem 4.3 that

$$h(\sqrt{n}(X_n - x^*))) \quad \xrightarrow[n\to+\infty]{\mathcal{L}} \quad \mathcal{N}(0, \sigma^2(\lambda)).$$

As the equilibrium point x^* and the asymptotic variance $\sigma^2(\lambda)$ are unknown, we use estimators from the standard Monte Carlo procedure. We denote for each fixed lambda $\widehat{\sigma}_n^2(\lambda)$ the estimator of $\sigma^2(\lambda)$. Given the form of our function h, we deduce that the limiting variances should be related as $(1-\lambda)^2\sigma^2(0) = \sigma^2(\lambda)$. The results are shown in Figure 2.



(a) SGD ($\lambda = 0$), where $\sigma_n^2(0) \approx 743.11$

(b) 0.5-SAGA, where $\sigma_n^2(0.5) \approx 5$

(c) 0.9-SAGA, where $\sigma_n^2(0.9) \approx 0.884$

(d) SAGA ($\lambda = 1$), where $\sigma_n^2(1) \approx 0.98$

Fig. 2: We used 1000 samples, where each one was obtained by running the associated algorithm for $n = 500000$ iterations.

The main purpose of this plot is to represent the decreasing behavior of the variance with respect to the parameter λ. Even though for the SAGA ($\lambda = 1$) we know that its variance converges to zero, for finite n we can only see that it is shrinking with respect to n to obtain at the limit a Dirac mass at 0. Here, the sample variances satisfy $\hat{\sigma}_n^2(0.9) < \hat{\sigma}_n^2(1)$. Nevertheless, they are still very close in the scale of the sample variance $\hat{\sigma}_n^2(0)$ of the SGD. We explain this as a consequence of the approximations and the fact that the models $\lambda = 0.9$ and $\lambda = 1$ are intimately related.

Finally, we present approximate results of the mean squared error $\mathbb{E}[\|X_n - x^*\|^2]$. For that purpose, we suppose that Assumption 4.5 is satisfied so that Theorem 4.6 holds. We run each algorithm for 100 epochs, where each epoch consists of 1000 iterations. In order to approximate the expectation, we apply the standard Monte Carlo procedure with 1000 samples. Here, the approximation of x^* is the result of running the SAGA algorithm for $40M$ iterations. The results are illustrated in Figure 3. The plot just gives an intuition on the behavior of the mean squared error, since the constant K in Theorem 4.6 is unknown.

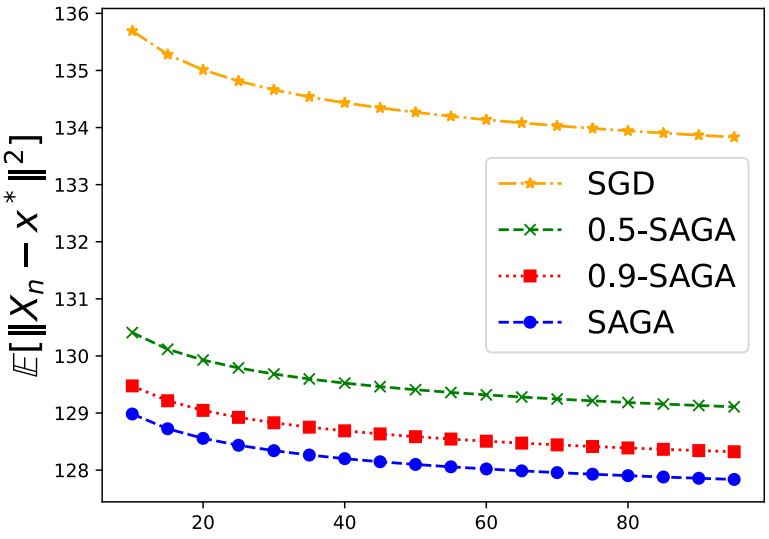


Fig. 3: Mean squared error with respect to epochs. We confirm the decreasing order of the mean squared error of $X_n - x^*$ with respect to λ and n.

6. Conclusion. Stochastic optimization is one of the main challenges of machine learning touching almost every aspect of the discipline. Thus, in order to meet expectations, the SGD algorithm has been studied at length. However, the advent of Big Data for model learning led to the development of more sophisticated stochastic methods. In our study, we therefore highlight the properties of the new λ-SAGA algorithm which is a generalization of the SAGA algorithm. We were able to establish the almost sure convergence and the asymptotic normality of this novel algorithm by using a decreasing step and without the strong convexity and Lipschitz gradient assumptions. The other major contribution of our paper concerns the convergence rates in $\mathbf{L}^p$ of the λ-SAGA algorithm. Finally, stochastic algorithms offer multiple guarantees in terms of convergence and certainly promise to continue to have profound impacts on the fast development of the machine learning field.

Appendix A. Some useful existing results. We first recall the well-known Robbins-Siegmund Theorem [34].

THEOREM A.1 (Robbins-Siegmund theorem). *Let $(V_n), (a_n), (\mathcal{A}_n), (\mathcal{B}_n)$ be four positive sequences of random variables adapted to a filtration $(\mathcal{F}_n)$ such that*

$$\mathbb{E}[V_{n+1}|\mathcal{F}_n] \leqslant (1 + a_n)V_n + \mathcal{A}_n - \mathcal{B}_n,$$

where

$$\sum_{n=1}^{\infty} a_n < +\infty \qquad \textit{and} \qquad \sum_{n=1}^{\infty} \mathcal{A}_n < +\infty \qquad a.s.$$

Then, (V_n) converges almost surely towards a finite random variable V and

$$\sum_{n=1}^{\infty} \mathcal{B}_n < +\infty \qquad a.s.$$

The next two lemma provide very useful inequality for non-asymptotic convergence rates. The first lemma is given by Lemma A.3 in supplementary material of [5], see also Theorem 1 in [4].

LEMMA A.2. *[5]. Let (Z_n) be a sequence of positive real numbers satisfying, for all $n \geqslant 1$, the recursive inequality*

$$Z_{n+1} \leqslant \left(1 - \frac{a}{(n+1)^\alpha}\right) Z_n + \frac{b}{(n+1)^\beta}, \tag{A.1}$$

where a, b, α and β are positive constants satisfying $a \leqslant 2^\alpha$, $\alpha \leqslant 1$, $1 < \beta < 2$ and $\beta \leqslant 2\alpha$ with $\beta < a + 1$ in the special case where $\alpha = 1$. Then, there exists a positive constant C such that, for any $n \geqslant 1$,

$$Z_n \leqslant \frac{C}{n^{\beta-\alpha}}. \tag{A.2}$$

The second lemma is given without proof in [10] in the special case $p \in (0, 2]$, see Lemma B.3 as well as the seminal paper [2]. We extend it to the case p even and we propose a short proof for the sake of completeness.

LEMMA A.3. *Let p be a positive even integer. There exist two positive constant C_p and D_p such that for any a, $b \in \mathbb{R}^d$,*

$$\|a + b\|^{2+p} \leqslant \|a\|^{2+p} + (2 + p)\langle a, b\rangle \|a\|^p + C_p \|a\|^p \|b\|^2 + D_p \|b\|^{2+p}. \tag{A.3}$$

Proof. We prove Lemma A.3 by induction. For the base case $p = 2$, we have

$$\begin{aligned} \|a + b\|^4 &= \left(\|a\|^2 + 2\langle a, b\rangle + \|b\|^2\right)^2 \\ &= \|a\|^4 + 4(\langle a, b\rangle)^2 + \|b\|^4 + 4\langle a, b\rangle \|a\|^2 + 4\langle a, b\rangle \|b\|^2 + 2\|a\|^2 \|b\|^2. \end{aligned}$$

It follows from Cauchy–Schwarz inequality that $(\langle a, b\rangle)^2 \leqslant \|a\|^2 \|b\|^2$. Moreover, we also have $2\langle a, b\rangle \leqslant \|a\|^2 + \|b\|^2$ which implies that $4\langle a, b\rangle \|b\|^2 \leqslant 2\|a\|^2 \|b\|^2 + 2\|b\|^4$. Hence, we obtain from these two inequalities that

$$\|a + b\|^4 \leqslant \|a\|^4 + 4\langle a, b\rangle \|a\|^2 + 8\|a\|^2 \|b\|^2 + 3\|b\|^4,$$

which leads to $C_2 = 8$ and $D_2 = 3$. Hereafter, assume that inequality (A.3) holds up to $q \geq 2$ and let $p = 2 + q$. We have by induction

$$\begin{aligned} &\|a + b\|^{2+p} = \|a + b\|^2 \|a + b\|^p = \|a + b\|^2 \|a + b\|^{2+q} \\ &\leqslant \left(\|a\|^2 + 2\langle a, b\rangle + \|b\|^2\right)\left(\|a\|^{2+q} + (2 + q)\langle a, b\rangle \|a\|^q + C_q \|a\|^q \|b\|^2 + D_q \|b\|^{2+q}\right) \\ &\leqslant \|a\|^{2+p} + (2 + p)\langle a, b\rangle \|a\|^p + (C_q + 2p + 1)\|a\|^p \|b\|^2 + (2C_q + p)\|a\|^{q+1} \|b\|^3 \\ &+ D_q \|a\|^2 \|b\|^p + C_q \|a\|^q \|b\|^4 + 2D_q \|a\| \|b\|^{p+1} + D_q \|b\|^{2+p} \end{aligned}$$

where the last inequality is the result of applying Cauchy-Schwarz inequality $\langle a, b\rangle \leq \|a\| \|b\|$ for all the terms multiplied by $\langle a, b\rangle$ but the ones isolated in the second term.

Furthermore, it follows from Young's inequality for products that

$$\|a\|^{q+1}\|b\|^3 = \|a\|^{q+1}\|b\|^{2(q+1)/p} \times \|b\|^{3-2(q+1)/p} \leq \frac{\|a\|^p\|b\|^2}{p/(p-1)} + \frac{\|b\|^{p+2}}{p},$$

$$\|a\|^2\|b\|^p = \|a\|^2\|b\|^{4/p} \times \|b\|^{p-4/p} \leq \frac{\|a\|^p\|b\|^2}{p/2} + \frac{\|b\|^{p+2}}{p/q},$$

$$\|a\|^q\|b\|^4 = \|a\|^q\|b\|^{2q/p} \times \|b\|^{4-2q/p} \leq \frac{\|a\|^p\|b\|^2}{p/q} + \frac{\|b\|^{p+2}}{p/2},$$

$$\|a\|\|b\|^{p+1} = \|a\|\|b\|^{2/p} \times \|b\|^{p+1-2/p} \leq \frac{\|a\|^p\|b\|^2}{p} + \frac{\|b\|^{p+2}}{p/(p-1)}.$$

Finally, we obtain (A.3) with C_p and D_p satisfying the system defined, for $p \geqslant 4$, by

$$\left\{ \begin{array}{rcl} C_p & = & 3p + \dfrac{4}{p}\Big((p-1)C_{p-2} + D_{p-2}\Big) \\ D_p & = & 1 + \dfrac{4}{p}\Big(C_{p-2} + (p-1)D_{p-2}\Big) \end{array} \right.$$

with initial values $C_2 = 8$ and $D_2 = 3$, which achieves the proof of Lemma A.3. ∎

Remark A.4. One can easily compute $C_4 = 39$ and $D_4 = 18$. Moreover, one can observe that we always have $D_p \leqslant C_p$. Consequently, we can make use of (A.3) with C_p instead of D_p.

Appendix B. Proof of Theorem 4.3.

Proof. The λ-SAGA algorithm can be rewritten as

$$X_{n+1} = X_n - \gamma_n(\nabla f(X_n) + \varepsilon_{n+1}),$$

where $\varepsilon_{n+1} = \mathcal{Y}_{n+1} - \lambda Z_{n+1}$ with

$$\left\{ \begin{array}{ll} \mathcal{Y}_{n+1} & = \nabla f_{U_{n+1}}(X_n) - \nabla f(X_n), \\ Z_{n+1} & = \nabla f_{U_{n+1}}(\phi_{n,U_{n+1}}) - \dfrac{1}{N}\displaystyle\sum_{k=1}^{N} \nabla f_k(\phi_{n,k}). \end{array} \right.$$

We already saw that (ε_n) is a martingale difference adapted to the filtration $(\mathcal{F}_n)$. Moreover,

$$\begin{aligned} \mathbb{E}[\varepsilon_{n+1}\varepsilon_{n+1}^T|\mathcal{F}_n] = \mathbb{E}[\mathcal{Y}_{n+1}\mathcal{Y}_{n+1}^T|\mathcal{F}_n] - \lambda\mathbb{E}[\mathcal{Y}_{n+1}Z_{n+1}^T|\mathcal{F}_n] - \lambda\mathbb{E}[Z_{n+1}\mathcal{Y}_{n+1}^T|\mathcal{F}_n] \\ + \lambda^2\mathbb{E}[Z_{n+1}Z_{n+1}^T|\mathcal{F}_n]. \end{aligned}$$

In addition, we clearly have that almost surely

$$\mathbb{E}[\mathcal{Y}_{n+1}\mathcal{Y}_{n+1}^T|\mathcal{F}_n] = \frac{1}{N}\sum_{k=1}^{N} \nabla f_k(X_n)\left(\nabla f_k(X_n)\right)^T - \nabla f(X_n)\left(\nabla f(X_n)\right)^T,$$

$$\mathbb{E}[Z_{n+1}Z_{n+1}^T|\mathcal{F}_n] = \frac{1}{N}\sum_{k=1}^{N} \nabla f_k(\phi_{n,k})\left(\nabla f_k(\phi_{n,k})\right)^T - \frac{1}{N^2}\left(\sum_{k=1}^{N} \nabla f_k(\phi_{n,k})\right)\left(\sum_{k=1}^{N} \left(\nabla f_k(\phi_{n,k})\right)^T\right),$$

$$\mathbb{E}[Z_{n+1}\mathcal{Y}_{n+1}^T|\mathcal{F}_n] = \frac{1}{N}\sum_{k=1}^{N} \nabla f_k(\phi_{n,k})\left(\nabla f_k(X_n)\right)^T - \frac{1}{N}\sum_{k=1}^{N} \nabla f_k(\phi_{n,k})\left(\nabla f(X_n)\right)^T.$$

∎

We now claim that for all $1 \leqslant k \leqslant N$

$$\lim_{n\to+\infty} \phi_{n,k} = x^* \qquad a.s. \tag{B.1}$$

As a matter of fact, for a fixed value $1 \leqslant k \leqslant N$, the probability that $U_n = k$ occurs for infinitely many n. Consequently, $(\phi_{n,k})$ is a sub-sequence of (X_n), since $\phi_{n,k}$ is updated to X_n each time $\{U_n = k\}$. Hence, the almost sure convergence (B.1) follows from (4.16). Combining the almost sure convergence of (X_n) and $(\phi_{n,k})$ towards x^* with the continuity of ∇f given by Assumption 4.2, it follows that almost surely

$$\begin{cases} \lim_{n\to+\infty} \mathbb{E}[\mathcal{Y}_{n+1}\mathcal{Y}_{n+1}^T|\mathcal{F}_n] &= \Gamma, \\ \lim_{n\to+\infty} \mathbb{E}[\mathcal{Y}_{n+1}Z_{n+1}^T|\mathcal{F}_n] &= \Gamma, \\ \lim_{n\to+\infty} \mathbb{E}[Z_{n+1}\mathcal{Y}_{n+1}^T|\mathcal{F}_n] &= \Gamma, \\ \lim_{n\to+\infty} \mathbb{E}[Z_{n+1}Z_{n+1}^T|\mathcal{F}_n] &= \Gamma, \end{cases}$$

where

$$\Gamma = \frac{1}{N}\sum_{k=1}^{N} \nabla f_k(x^*)\left(\nabla f_k(x^*)\right)^T,$$

which leads to

$$\lim_{n\to+\infty} \mathbb{E}[\varepsilon_{n+1}\varepsilon_{n+1}^T|\mathcal{F}_n] = (1-\lambda)^2\Gamma \qquad a.s.$$

Therefore, we obtain from Toeplitz's lemma that

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[\varepsilon_k\varepsilon_k^T|\mathcal{F}_{k-1}] = (1-\lambda)^2\Gamma \qquad a.s.$$

In addition, we have for all $n \geqslant 1$

$$\|\varepsilon_{n+1}\| \leqslant 2\left(\max_{k=1,\dots,N}\|\nabla f_k(X_n)\| + \lambda \max_{k=1,\dots,N}\|\nabla f_k(\phi_{n,k})\|\right). \tag{B.2}$$

Hence, it follows from (B.2) that

$$\|\varepsilon_{n+1}\|^4 \leqslant 128\left(\max_{k=1,\dots,N}\|\nabla f_k(X_n)\|^4 + \lambda^4 \max_{k=1,\dots,N}\|\nabla f_k(\phi_{n,k})\|^4\right). \tag{B.3}$$

However, for all $k = 1, ..., N$, we have

$$\|\nabla f_k(X_n)\|^2 \leqslant 2\|\nabla f_k(X_n) - \nabla f_k(x^*)\|^2 + 2\|\nabla f_k(x^*)\|^2,$$

which implies that

$$\max_{k=1,\dots,N}\|\nabla f_k(X_n)\|^2 \leqslant 2N\Big(\tau^2(X_n) + \theta^*\Big).$$

Consequently,

$$\max_{k=1,\dots,N}\|\nabla f_k(X_n)\|^4 \leqslant 4N^2\Big(\tau^2(X_n) + \theta^*\Big)^2. \tag{B.4}$$

By the same token,

$$\max_{k=1,\ldots,N} \|\nabla f_k(\phi_{n,k})\|^4 \leqslant 4N^2 \Big(A_n + \theta^*\Big)^2. \tag{B.5}$$

Hence, we obtain from (B.3), (B.4) and (B.5) that

$$\|\varepsilon_{n+1}\|^4 \leqslant 512N^2 \left(\Big(\tau^2(X_n) + \theta^*\Big)^2 + \lambda^4 \Big(A_n + \theta^*\Big)^2 \right),$$

which immediately implies that

$$\mathbb{E}[\|\varepsilon_{n+1}\|^4 | \mathcal{F}_n] \leqslant 512N^2 \left(\Big(\tau^2(X_n) + \theta^*\Big)^2 + \lambda^4 \Big(A_n + \theta^*\Big)^2 \right) \qquad a.s. \tag{B.6}$$

Moreover, since X_n converges towards x^*, it follows that $\tau^2(X_n)$ converges to 0 almost surely. Combining this result with the almost sure convergence of A_n towards 0 and (B.6), we find that

$$\sup_{n \geqslant 1} \mathbb{E}[\|\varepsilon_{n+1}\|^4 | \mathcal{F}_n] < +\infty,$$

which implies that for all $\epsilon > 0$,

$$\lim_{n \to +\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}\left[\|\varepsilon_k\|^2 \mathbb{1}_{\{\|\varepsilon_k\| \geqslant \epsilon \sqrt{n}\}} | \mathcal{F}_{k-1} \right] = 0 \qquad a.s.$$

Finally, it follows from the central limit theorem for stochastic algorithms given by Theorem 2.3 in [41] that

$$\sqrt{n}(X_n - x^*) \quad \xrightarrow[n \to +\infty]{\mathcal{L}} \quad \mathcal{N}_d(0, \Sigma),$$

where

$$\Sigma = (1-\lambda)^2 \int_0^\infty (e^{-(H - \mathbb{I}_d/2)u})^T \Gamma e^{-(H - \mathbb{I}_d/2)u} du,$$

which completes the proof of Theorem 4.3. ∎

Appendix C. Proof of Theorem 4.6.

Proof. We already saw in (4.10) that for all $n \geq 1$,

$$\mathbb{E}[V_{n+1} | \mathcal{F}_n] \leqslant V_n - 2\gamma_n \langle X_n - x^*, \nabla f(X_n) \rangle + 2\gamma_n^2 \big(\tau^2(X_n) + A_n + \theta^*\big) \qquad a.s.$$

Hence, it follows from Assumption 4.5 that $\langle X_n - x^*, \nabla f(X_n) \rangle \geqslant \mu V_n$, which leads to

$$\mathbb{E}[V_{n+1} | \mathcal{F}_n] \leqslant (1 - 2\mu\gamma_n) V_n + 2\gamma_n^2 \big(\tau^2(X_n) + A_n + \theta^*\big) \qquad a.s.$$

By taking the expectation on both side of this inequality, we obtain that for all $n \geq 1$,

$$\mathbb{E}[V_{n+1}] \leqslant (1 - 2\mu\gamma_n) \mathbb{E}[V_n] + 2\gamma_n^2 \big(\mathbb{E}[\tau^2(X_n)] + \mathbb{E}[A_n] + \theta^*\big). \tag{C.1}$$

Furthermore, we deduce from Corollary E.2 in Appendix E below that there exist positive constants b_1 and b_2 such that, for all $n \geqslant 1$, $\mathbb{E}[\tau^2(X_n)] \leqslant b_1$ and $\mathbb{E}[A_n] \leqslant b_2$. Consequently, (C.1) immediately leads, for all $n \geqslant 1$, to

$$\mathbb{E}[V_{n+1}] \leqslant \left(1 - \frac{a}{(n+1)^\alpha}\right) \mathbb{E}[V_n] + \frac{b}{(n+1)^{2\alpha}}$$

where $a = 2\mu c$ and $b = c^2 2^{2\alpha+1}(b_1 + b_2 + \theta^*)$. Finally, we can conclude from Lemma A.2 that there exists a positive constant K such that for any $n \geqslant 1$,

$$\mathbb{E}[\|X_n - x^*\|^2] \leqslant \frac{K}{n^\alpha},$$

which completes the proof of Theorem 4.6. □

Appendix D. Proof of Theorem 4.8.

Proof. First of all, Theorem 4.8 follows from Theorem 4.6 in the special case $p = 1$. Hence, we are going to prove Theorem 4.8 by induction on $n \geqslant 1$ for some integer $p \geqslant 2$ satisfying Assumption 4.7. As the initial state X_1 belongs to $\mathbf{L}^{2p}$, the base case is immediately true. Next, assume by induction that for some integer m which will be fixed soon, there exists a positive constant K_p such that for all $n \leqslant m$,

$$\mathbb{E}[\|X_n - x^*\|^{2p}] \leqslant \frac{K_p}{n^{p\alpha}}. \tag{D.1}$$

We have from (4.3) together with Lemma A.3 that it exists a positive constant C_p such that for all $n \geqslant 1$,

$$\begin{aligned} V_{n+1}^p &= \|X_{n+1} - x^*\|^{2p}, \\ &= \|X_n - x^* - \gamma_n(Y_{n+1} - \lambda Z_{n+1})\|^{2p}, \\ &\leqslant V_n^p - 2p\gamma_n V_n^{p-1}\langle X_n - x^*, Y_{n+1} - \lambda Z_{n+1}\rangle + C_p\gamma_n^2 V_n^{p-1}\|Y_{n+1} - \lambda Z_{n+1}\|^2 \\ &\quad + C_p\gamma_n^{2p}\|Y_{n+1} - \lambda Z_{n+1}\|^{2p}. \end{aligned}$$

Hence, it follows from (4.4) that

$$\begin{aligned} \mathbb{E}[V_{n+1}^p|\mathcal{F}_n] &\leqslant V_n^p - 2p\gamma_n V_n^{p-1}\langle X_n - x^*, \nabla f(X_n)\rangle \\ &\quad + C_p\gamma_n^2 V_n^{p-1}\mathbb{E}[\|Y_{n+1} - \lambda Z_{n+1}\|^2|\mathcal{F}_n] + C_p\gamma_n^{2p}\mathbb{E}[\|Y_{n+1} - \lambda Z_{n+1}\|^{2p}|\mathcal{F}_n]. \end{aligned} \tag{D.2}$$

We already saw from (4.9) that

$$\mathbb{E}[\|Y_{n+1} - \lambda Z_{n+1}\|^2|\mathcal{F}_n] \leqslant 2\big(\tau^2(X_n) + A_n + \theta^*\big) \qquad a.s.$$

which leads, via Assumption 4.7, to

$$\mathbb{E}[\|Y_{n+1} - \lambda Z_{n+1}\|^2|\mathcal{F}_n] \leqslant 2\big(L_p^{1/p}\, V_n + A_n + \theta^*\big) \qquad a.s.$$

Moreover, as in the proof of (4.5), we deduce from Jensen's inequality that

$$\begin{aligned} \mathbb{E}[\|Y_{n+1} - \lambda Z_{n+1}\|^{2p}|\mathcal{F}_n] &\leqslant 3^{2p-1}\mathbb{E}[\|Y_{n+1} - \nabla f_{U_{n+1}}(x^*)\|^{2p}|\mathcal{F}_n] \\ &\quad + 3^{2p-1}\mathbb{E}[\|Z_{n+1} - \nabla f_{U_{n+1}}(x^*)\|^{2p}|\mathcal{F}_n] + 3^{2p-1}\mathbb{E}[\|\nabla f_{U_{n+1}}(x^*)\|^{2p}|\mathcal{F}_n]. \end{aligned} \tag{D.3}$$

Hereafter, we clearly have

$$\mathbb{E}[\|\nabla f_{U_{n+1}}(x^*)\|^{2p}|\mathcal{F}_n] = \frac{1}{N}\sum_{k=1}^N \|\nabla f_k(x^*)\|^{2p} \qquad a.s. \tag{D.4}$$

Moreover, denote

$$A_{p,n} = \frac{1}{N}\sum_{k=1}^N \|\nabla f_k(\phi_{n,k}) - \nabla f_k(x^*)\|^{2p} \qquad \text{and} \qquad \Sigma_n = \frac{1}{N}\sum_{k=1}^N \nabla f_k(\phi_{n,k}).$$

It follows once again from Jensen's inequality that

$$\mathbb{E}[\|Z_{n+1} - \nabla f_{U_{n+1}}(x^*)\|^{2p}|\mathcal{F}_n] \leqslant 2^{2p-1}\big(A_{p,n} + \|\Sigma_n\|^{2p}\big) \qquad a.s. \tag{D.5}$$

However, we obtain from Holder's inequality that $\|\Sigma_n\|^{2p} \leqslant A_{p,n}$. Consequently, inequality (D.5) immediately leads to

$$\mathbb{E}[\|Z_{n+1} - \nabla f_{U_{n+1}}(x^*)\|^{2p}|\mathcal{F}_n] \leqslant 4^p A_{p,n} \qquad a.s. \tag{D.6}$$

Furthermore, define for all $x \in \mathbb{R}^d$,

$$\tau^{2p}(x) = \frac{1}{N}\sum_{k=1}^{N}\|\nabla f_k(x) - \nabla f_k(x^*)\|^{2p}.$$

As in the proof of Theorem 4.1, one can observe that

$$\mathbb{E}[\|Y_{n+1} - \nabla f_{U_{n+1}}(x^*)\|^{2p}|\mathcal{F}_n] = \tau^{2p}(X_n) \qquad a.s. \tag{D.7}$$

Putting together the three contributions (D.4), (D.6) and (D.7), we obtain from (D.3) that

$$\mathbb{E}[\|Y_{n+1} - \lambda Z_{n+1}\|^{2p}|\mathcal{F}_n] \leqslant 3^{2p-1}(\tau^{2p}(X_n) + 4^p A_{p,n} + \theta_p^*) \qquad a.s. \tag{D.8}$$

where

$$\theta_p^* = \frac{1}{N}\sum_{k=1}^{N}\|\nabla f_k(x^*)\|^{2p}.$$

Hence, Assumption 4.7 implies that

$$\mathbb{E}[\|Y_{n+1} - \lambda Z_{n+1}\|^{2p}|\mathcal{F}_n] \leqslant 3^{2p-1}(L_p V_n^p + 4^p A_{p,n} + \theta_p^*) \qquad a.s. \tag{D.9}$$

Therefore, we deduce from (D.2) and (D.9) that for all $n \geqslant 1$,

$$\begin{aligned}\mathbb{E}[V_{n+1}^p|\mathcal{F}_n] &\leqslant \Big(1 + 2L_p^{1/p}\, C_p\gamma_n^2 + 3^{2p-1}L_p C_p \gamma_n^{2p}\Big)V_n^p \\ &\quad - 2p\gamma_n V_n^{p-1}\langle X_n - x^*, \nabla f(X_n)\rangle + 2C_p\gamma_n^2 V_n^{p-1}(A_n + \theta^*) \\ &\quad + 3^{2p-1}C_p\gamma_n^{2p}(4^p A_{p,n} + \theta_p^*) \qquad a.s.\end{aligned}$$

Furthermore, it follows from Assumption 4.5 that $\langle X_n - x^*, \nabla f(X_n)\rangle \geqslant \mu V_n$, which leads to

$$\begin{aligned}\mathbb{E}[V_{n+1}^p|\mathcal{F}_n] &\leqslant \Big(1 + 2L_p^{1/p}\, C_p\gamma_n^2 + 3^{2p-1}L_p C_p \gamma_n^{2p} - 2\mu p\gamma_n\Big)V_n^p \\ &\quad + 2C_p\gamma_n^2 V_n^{p-1}(A_n + \theta^*) + 3^{2p-1}C_p\gamma_n^{2p}(4^p A_{p,n} + \theta_p^*) \qquad a.s.\end{aligned} \tag{D.10}$$

By taking the expectation on both side of this inequality, we obtain that for all $n \geq 1$,

$$\begin{aligned}\mathbb{E}[V_{n+1}^p] &\leqslant \Big(1 + 2L_p^{1/p}\, C_p\gamma_n^2 + 3^{2p-1}L_p C_p \gamma_n^{2p} - 2\mu p\gamma_n\Big)\mathbb{E}[V_n^p] + 2C_p\gamma_n^2\mathbb{E}[A_n V_n^{p-1}] \\ &\quad + 2C_p\gamma_n^2\theta^*\mathbb{E}[V_n^{p-1}] + 3^{2p-1}C_p\gamma_n^{2p}(4^p\mathbb{E}[A_{p,n}] + \theta_p^*)\end{aligned} \tag{D.11}$$

We deduce from Corollary F.2 in Appendix F below that there exists a positive constant d_p such that for all $n \geqslant 1$, $\mathbb{E}[A_{p,n}] \leqslant d_p$. The main difficulty arising here is

to find a sharp upper bound for the crossing term $\mathbb{E}[A_n V_n^{p-1}]$. By using once again Holder's inequality, we have for all $n \geqslant 1$,

$$\mathbb{E}[A_n V_n^{p-1}] \leq \Big(\mathbb{E}[A_n^p]\Big)^{\frac{1}{p}} \Big(\mathbb{E}[V_n^p]\Big)^{\frac{p-1}{p}} \qquad \text{and} \qquad \mathbb{E}[V_n^{p-1}] \leq \Big(\mathbb{E}[V_n^p]\Big)^{\frac{p-1}{p}}. \tag{D.12}$$

Hence, it is necessary to compute an upper bound for $\mathbb{E}[A_n^p]$. Nevertheless, one can observe from Jensen's inequality that we always have $A_n^p \leqslant A_{p,n}$, which leads that $\mathbb{E}[A_n^p] \leqslant d_p$. Therefore, it follows from the induction hypothesis (D.1) together with (D.11) and (D.12) that

$$\mathbb{E}[V_{n+1}^p] \leqslant \Big(1 + \xi_n - 2\mu p \gamma_n\Big)\mathbb{E}[V_n^p] + \frac{b}{(n+1)^{\alpha(p+1)}} \tag{D.13}$$

where $\xi_n = 2L_p^{1/p} C_p \gamma_n^2 + 3^{2p-1} L_p C_p \gamma_n^{2p}$ and

$$b = 2^{\alpha(p+1)} c^2 C_p \Big(2K_p^{(p-1)/p}\big(d_p^{1/p} + \theta^*\big) + c^{p-1} 3^{2p-1}\big(4^p d_p + \theta_p^*\big)\Big).$$

Hereafter, denote by m the integer part of the real number

$$\Big(\frac{cC_p}{\mu p}\Big)^{1/\alpha} \Big(2L_p^{1/p} + 3^{2p-1} L_p\Big)^{1/\alpha}.$$

One can easily check that as soon as $n \geqslant m$, $\xi_n \leq \mu p \gamma_n$. Consequently, we find from (D.13) that as soon as $n \geqslant m$,

$$\mathbb{E}[V_{n+1}^p] \leqslant \left(1 - \frac{a}{(n+1)^\alpha}\right) \mathbb{E}[V_n^p] + \frac{b}{(n+1)^{\alpha(p+1)}} \tag{D.14}$$

where $a = p\mu c$. Finally, we deduce from Lemma A.2 that there exists a positive constant K_p such that for all $n \geqslant 1$,

$$\mathbb{E}[\|X_{n+1} - x^*\|^{2p}] \leqslant \frac{K_p}{(n+1)^{\alpha p}},$$

which achieves the induction on n and completes the proof of Theorem 4.8. ∎

Appendix E. Additional asymptotic result on the convergence in $\mathbf{L}^2$. The goal of this appendix is to provide additional asymptotic properties of the λ-SAGA algorithm that will be useful in the proofs of our main results. First of all, we recall that $V_n = \|X_n - x^*\|^2$,

$$A_n = \frac{1}{N}\sum_{k=1}^N \|\nabla f_k(\phi_{n,k}) - \nabla f_k(x^*)\|^2 \quad \text{and} \quad \tau^2(x) = \frac{1}{N}\sum_{k=1}^N \|\nabla f_k(x) - \nabla f_k(x^*)\|^2.$$

THEOREM E.1. *Consider a fixed $\lambda \in [0,1]$. Let (X_n) be the sequence generated by the λ-SAGA algorithm with decreasing step sequence (γ_n) satisfying (1.2). Suppose that Assumptions 3.1, 3.3 and 4.5 are satisfied. Then, we have almost surely*

$$\sum_{n=1}^\infty \gamma_n V_n < +\infty, \qquad \sum_{n=1}^\infty \gamma_n A_n < +\infty, \qquad \sum_{n=1}^\infty \gamma_n \tau^2(X_n) < +\infty. \tag{E.1}$$

In addition, we also have

$$\sum_{n=1}^\infty \gamma_n \mathbb{E}[V_n] < +\infty, \qquad \sum_{n=1}^\infty \gamma_n \mathbb{E}[A_n] < +\infty, \qquad \sum_{n=1}^\infty \gamma_n \mathbb{E}[\tau^2(X_n)] < +\infty. \tag{E.2}$$

Proof. Let us consider the same Lyapounov function used in the proof of Theorem 4.1 and given by (4.11). We recall from inequality (4.14) that for all $n \geqslant 1$,

$$\mathbb{E}[T_{n+1}|\mathcal{F}_n] \leqslant (1 + 4L\gamma_n^2)T_n - 2\gamma_n\langle X_n - x^*, \nabla f(X_n)\rangle + 2\theta^*\gamma_n^2 \qquad a.s. \tag{E.3}$$

Let $(\mathcal{T}_n)$ be the sequence of Lyapunov functions defined, for all $n \geqslant 2$, by $\mathcal{T}_n = b_n T_n$ where

$$b_n = \prod_{k=1}^{n-1} \left(1 + 4L\gamma_k^2\right)^{-1}.$$

Since $b_n = b_{n+1}(1 + 4L\gamma_n^2)$, we obtain from (E.3) that for all $n \geqslant 1$,

$$\mathbb{E}[\mathcal{T}_{n+1}|\mathcal{F}_n] \leqslant \mathcal{T}_n - 2\gamma_n b_{n+1}\langle X_n - x^*, \nabla f(X_n)\rangle + 2\theta^* b_{n+1}\gamma_n^2 \qquad a.s.$$

Hence, it follows from Assumption 4.5 that for all $n \geqslant 1$,

$$\mathbb{E}[\mathcal{T}_{n+1}|\mathcal{F}_n] \leqslant \mathcal{T}_n - 2\mu\gamma_n b_{n+1}V_n + 2\theta^* b_{n+1}\gamma_n^2 \qquad a.s. \tag{E.4}$$

Moreover, we clearly have from the right-hand side of (1.2) that (b_n) converges to a positive real number b, which implies that

$$\sum_{n=1}^{\infty} b_{n+1}\gamma_n^2 < +\infty. \tag{E.5}$$

Therefore, we deduce from the Robbins-Siegmund Theorem [34] given by Theorem A.1 that $(\mathcal{T}_n)$ converges almost surely towards a finite random variable $\mathcal{T}$ and the series

$$\sum_{n=2}^{\infty} \gamma_n b_{n+1} V_n < +\infty \qquad a.s.$$

which leads to

$$\sum_{n=1}^{\infty} \gamma_n V_n < +\infty \qquad a.s. \tag{E.6}$$

We also obtain from relation (E.6) and Assumption 3.3 that

$$\sum_{n=1}^{\infty} \gamma_n \tau^2(X_n) < +\infty \qquad a.s. \tag{E.7}$$

In addition, by taking the expectation of both sides of (E.4) and using a standard telescoping argument, we find that

$$2\mu \sum_{n=2}^{\infty} \gamma_n b_{n+1}\mathbb{E}[V_n] \leqslant \mathbb{E}[\mathcal{T}_2] + 2\theta^* \sum_{n=2}^{\infty} b_{n+1}\gamma_n^2. \tag{E.8}$$

Then, it follows from (E.5) and (E.8) that

$$\sum_{n=2}^{\infty} \gamma_n b_{n+1}\mathbb{E}[V_n] < +\infty,$$

which implies that

$$\sum_{n=1}^{\infty} \gamma_n \mathbb{E}[V_n] < +\infty.$$

Consequently, we get from Assumption 3.3 that

$$\sum_{n=1}^{\infty} \gamma_n \mathbb{E}[\tau^2(X_n)] < +\infty. \tag{E.9}$$

Furthermore, we already saw from (4.12) that for all $n \geqslant 1$,

$$\mathbb{E}[A_{n+1}|\mathcal{F}_n] = \frac{1}{N}\tau^2(X_n) + \left(1 - \frac{1}{N}\right) A_n \qquad a.s. \tag{E.10}$$

For all $n \geqslant 1$, denote $\mathcal{A}_n = \gamma_n A_n$. Since $\gamma_{n+1} \leqslant \gamma_n$, we obtain from (E.10) that for all $n \geqslant 1$,

$$\mathbb{E}[\mathcal{A}_{n+1}|\mathcal{F}_n] \leqslant \mathcal{A}_n + \frac{1}{N}\gamma_n \tau^2(X_n) - \frac{1}{N}\gamma_n A_n \qquad a.s. \tag{E.11}$$

By considering the almost sure convergence (E.7), it follows once again from the Robbins-Siegmund Theorem [34] given by Theorem A.1 that $(\mathcal{A}_n)$ converges almost surely towards a finite random variable $\mathcal{A}$ and the series

$$\sum_{n=1}^{\infty} \gamma_n A_n < +\infty \qquad a.s.$$

Moreover, by taking the expectation of both sides of (E.11) and using a standard telescoping argument, we obtain that

$$\sum_{n=1}^{\infty} \gamma_n \mathbb{E}[A_n] \leqslant N\mathbb{E}[\mathcal{A}_1] + \sum_{n=1}^{\infty} \gamma_n \mathbb{E}[\tau^2(X_n)]. \tag{E.12}$$

Finally, we deduce from (E.9) and (E.12) that

$$\sum_{n=1}^{\infty} \gamma_n \mathbb{E}[A_n] < +\infty,$$

which completes the proof of Theorem E.1. ∎

A straightforward application of Theorem E.1, using the left-hand side of (1.2), is as follows.

Corollary E.2. *Assume that the conditions of Theorem E.1 hold. Then, we have*

$$\lim_{n\to+\infty} V_n = \lim_{n\to+\infty} A_n = \lim_{n\to+\infty} \tau^2(X_n) = 0 \qquad a.s.$$

Moreover, we also have

$$\lim_{n\to+\infty} \mathbb{E}[V_n] = \lim_{n\to+\infty} \mathbb{E}[A_n] = \lim_{n\to+\infty} \mathbb{E}[\tau^2(X_n)] = 0.$$

Appendix F. Additional asymptotic result on the convergence in $\mathbf{L}^p$. As in the previous Appendix, our purpose is to establish additional asymptotic properties of the λ-SAGA algorithm that will be useful in the proofs of our main results. First of all, we recall that $V_n^p = \|X_n - x^*\|^{2p}$,

$$A_{p,n} = \frac{1}{N}\sum_{k=1}^{N} \|\nabla f_k(\phi_{n,k}) - \nabla f_k(x^*)\|^{2p} \quad \text{and} \quad \tau^{2p}(x) = \frac{1}{N}\sum_{k=1}^{N} \|\nabla f_k(x) - \nabla f_k(x^*)\|^{2p}.$$

THEOREM F.1. *Consider a fixed $\lambda \in [0,1]$. Let (X_n) be the sequence generated by the λ-SAGA algorithm with decreasing step γ_n satisfying (1.2). Suppose that Assumptions 3.1, 4.5 and 4.7 are satisfied. Then, we have almost surely*

$$\sum_{n=1}^{\infty} \gamma_n V_n^p < +\infty, \qquad \sum_{n=1}^{\infty} \gamma_n A_{p,n} < +\infty, \qquad \sum_{n=1}^{\infty} \gamma_n \tau^{2p}(X_n) < +\infty. \tag{F.1}$$

In addition, we also have

$$\sum_{n=1}^{\infty} \gamma_n \mathbb{E}[V_n^p] < +\infty, \qquad \sum_{n=1}^{\infty} \gamma_n \mathbb{E}[A_{p,n}] < +\infty, \qquad \sum_{n=1}^{\infty} \gamma_n \mathbb{E}[\tau^{2p}(X_n)] < +\infty. \tag{F.2}$$

Proof. We are going to prove Theorem F.1 by induction on $p \geqslant 1$. First of all, Theorem F.1 follows from Theorem E.1 in the special case $p = 1$. Hence, the base case is immediately true. Next, assume by induction that Theorem F.1 holds for some integer $p-1$ with $p \geqslant 2$. We recall from inequality (D.10) in the proof of Theorem 4.8 that for all $n \geqslant 1$,

$$\begin{aligned} \mathbb{E}[V_{n+1}^p|\mathcal{F}_n] &\leqslant \Big(1 + 2L_p^{1/p} C_p \gamma_n^2 + 3^{2p-1} L_p C_p \gamma_n^{2p} - 2\mu p \gamma_n\Big) V_n^p \\ &\quad + 2C_p \gamma_n^2 V_n^{p-1}(A_n + \theta^*) + 3^{2p-1} C_p \gamma_n^{2p}(4^p A_{p,n} + \theta_p^*) \qquad a.s. \end{aligned} \tag{F.3}$$

However, it follows from Young's inequality for products that almost surely

$$V_n^{p-1} A_n \leqslant \frac{A_n^p}{p} + \frac{(p-1)V_n^p}{p}. \tag{F.4}$$

Moreover, one can observe from Jensen's inequality that $A_n^p \leqslant A_{p,n}$ almost surely. Combining the previous inequality with (F.4), it implies that

$$V_n^{p-1} A_n \leqslant \frac{A_{p,n}}{p} + \frac{(p-1)V_n^p}{p}. \tag{F.5}$$

Furthermore, by putting together the inequalities (F.3) and (F.5), we obtain that

$$\begin{aligned} \mathbb{E}[V_{n+1}^p|\mathcal{F}_n] &\leqslant \Big(1 + 2\big(L_p^{1/p} + p^{-1}(p-1)\big) C_p \gamma_n^2 + 3^{2p-1} L_p C_p \gamma_n^{2p}\Big) V_n^p - 2\mu p \gamma_n V_n^p \\ &\quad + 2C_p \theta^* \gamma_n^2 V_n^{p-1} + \Big(3^{2p-1} 4^p + 2p^{-1}\Big) C_p \gamma_n^2 A_{p,n} + 3^{2p-1} C_p \gamma_n^{2p} \theta_p^* \qquad a.s. \end{aligned} \tag{F.6}$$

Let $(T_{p,n})$ be the sequence of Lyapunov functions defined, for all $n \geqslant 2$, by

$$T_{p,n} = V_n^p + N e_p \gamma_{n-1}^2 A_{p,n}, \tag{F.7}$$

where $e_p = C_p\Big(3^{2p-1}4^p + 2p^{-1}\Big)$. By the definition (F.7), we have

$$\mathbb{E}[T_{p,n+1}|\mathcal{F}_n] = \mathbb{E}[V_{n+1}^p|\mathcal{F}_n] + Ne_p\gamma_n^2\mathbb{E}[A_{p,n+1}|\mathcal{F}_n] \qquad a.s. \tag{F.8}$$

However, we deduce by the same arguments as in (4.12) that

$$\mathbb{E}[A_{p,n+1}|\mathcal{F}_n] = \frac{1}{N}\tau^{2p}(X_n) + \left(1 - \frac{1}{N}\right) A_{p,n} \qquad a.s. \tag{F.9}$$

Hence, it follows from (F.6), (F.8) and (F.9) that

$$\begin{aligned}\mathbb{E}[T_{p,n+1}|\mathcal{F}_n] &\leqslant T_{p,n} + C_p\left(2\big(L_p^{1/p} + p^{-1}(p-1)\big)\gamma_n^2 + 3^{2p-1}L_p\gamma_n^{2p}\right)V_n^p - 2\mu p\gamma_n V_n^p \\ &+ e_p\gamma_n^2\tau^{2p}(X_n) + 2C_p\theta^*\gamma_n^2 V_n^{p-1} + 3^{2p-1}C_p\gamma_n^{2p}\theta_p^* \qquad a.s.\end{aligned} \tag{F.10}$$

Additionally, we clearly have $V_n^p \leqslant T_{p,n}$ and Assumption 4.7 leads to

$$\tau^{2p}(X_n) \leqslant L_p V_n^p \leqslant L_p T_{p,n}.$$

Finally, we obtain from (F.10) that

$$\mathbb{E}[T_{p,n+1}|\mathcal{F}_n] \leqslant (1 + a_n)T_{p,n} + \Delta_n - 2\mu p\gamma_n V_n^p \qquad a.s., \tag{F.11}$$

where $\Delta_n = 2C_p\theta^*\gamma_n^2 V_n^{p-1} + 3^{2p-1}C_p\gamma_n^{2p}\theta_p^*$ and

$$a_n = e_p L_p\gamma_n^2 + 2C_p\big(L_p^{1/p} + p^{-1}(p-1)\big)\gamma_n^2 + 3^{2p-1}L_pC_p\gamma_n^{2p}.$$

Since the sequence (γ_n) satisfies (1.2), it is easy to see that

$$\sum_{n=1}^{\infty} a_n < +\infty.$$

Moreover, by the induction hypothesis, we have that

$$\begin{cases} \displaystyle\sum_{n=1}^{\infty}\gamma_n V_n^{p-1} < +\infty \qquad a.s., \\ \displaystyle\sum_{n=1}^{\infty}\gamma_n\mathbb{E}[V_n^{p-1}] < +\infty. \end{cases} \tag{F.12}$$

From (F.12), one can immediately deduce that

$$\begin{cases} \displaystyle\sum_{n=1}^{\infty}\Delta_n < +\infty \qquad a.s., \\ \displaystyle\sum_{n=1}^{\infty}\mathbb{E}[\Delta_n] < +\infty. \end{cases} \tag{F.13}$$

Therefore, one uses exactly the same lines as in the proof of Theorem E.1 and the Robbins-Siegmund Theorem [34] given by Theorem A.1 to show that

$$\begin{cases} \displaystyle\sum_{n=1}^{\infty}\gamma_n V_n^{p} < +\infty \qquad a.s., \\ \displaystyle\sum_{n=1}^{\infty}\gamma_n\mathbb{E}[V_n^{p}] < +\infty. \end{cases} \tag{F.14}$$

Hence, combining (F.14) with Assumption 4.7, one immediately deduces that

$$\text{(F.15)} \qquad \begin{cases} \displaystyle\sum_{n=1}^{\infty} \gamma_n \tau^{2p}(X_n) < +\infty \qquad a.s., \\ \displaystyle\sum_{n=1}^{\infty} \gamma_n \mathbb{E}[\tau^{2p}(X_n)] < +\infty. \end{cases}$$

Finally, using once again the same arguments as in the proof of Theorem E.1, we obtain that

$$\text{(F.16)} \qquad \begin{cases} \displaystyle\sum_{n=1}^{\infty} \gamma_n A_{p,n} < +\infty \qquad a.s., \\ \displaystyle\sum_{n=1}^{\infty} \gamma_n \mathbb{E}[A_{p,n}] < +\infty. \end{cases},$$

which achieves the proof of Theorem F.1. ∎

A useful consequence of Theorem F.1, using the left-hand side of (1.2), is as follows.

COROLLARY F.2. *Assume that the conditions of Theorem F.1 hold. Then, for all $p \geqslant 1$, we have*

$$\lim_{n\to+\infty} V_n^p = \lim_{n\to+\infty} A_{p,n} = \lim_{n\to+\infty} \tau^{2p}(X_n) = 0 \qquad a.s.,$$

and

$$\lim_{n\to+\infty} \mathbb{E}[V_n^p] = \lim_{n\to+\infty} \mathbb{E}[A_{p,n}] = \lim_{n\to+\infty} \mathbb{E}[\tau^{2p}(X_n)] = 0.$$

REFERENCES

[1] Z. Allen-Zhu, Katyusha: The first direct acceleration of stochastic gradient methods, Journal of Machine Learning Research, 18 (2018), pp. 1–51.

[2] P. Assouad, Espaces p-lisses et q-convexes. Inégalités de Burkholder, Séminaire Maurey-Schwartz, (1975), pp. 1–7. talk:15.

[3] F. Bach, Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression, The Journal of Machine Learning Research, 15 (2014), pp. 595–627.

[4] F. Bach and E. Moulines, Non-asymptotic analysis of stochastic approximation algorithms for machine learning, Advances in neural information processing systems, 24 (2011).

[5] B. Bercu and J. Bigot, Asymptotic distribution and convergence rates of stochastic algorithms for entropic optimal transportation between probability measures, The Annals of Statistics, 49 (2021), pp. 968–987.

[6] B. Bercu, A. Godichon, and B. Portier, An efficient stochastic Newton algorithm for parameter estimation in logistic regressions, SIAM J. Control Optim., 58 (2020), pp. 348–367.

[7] D. P. Bertsekas and J. N. Tsitsiklis, Gradient convergence in gradient methods with errors, SIAM Journal on Optimization, 10 (2000), pp. 627–642.

[8] L. Bottou, F. E. Curtis, and J. Nocedal, Optimization methods for large-scale machine learning, SIAM review, 60 (2018), pp. 223–311.

[9] N. Chatterji, N. Flammarion, Y. Ma, P. Bartlett, and M. Jordan, On the theory of variance reduction for stochastic gradient monte carlo, in International Conference on Machine Learning, PMLR, 2018, pp. 764–773.

[10] X. Chen, Z. Lai, H. Li, and Y. Zhang, Online statistical inference for stochastic optimization via kiefer-wolfowitz methods, Journal of the American Statistical Association, (2024), pp. 1–24.

[11] A. Defazio, A simple practical accelerated method for finite sums, Advances in neural information processing systems, 29 (2016).

[12] A. Defazio, F. Bach, and S. Lacoste-Julien, Saga: A fast incremental gradient method with support for non-strongly convex composite objectives, Advances in neural information processing systems, 27 (2014).
[13] M. Duflo, Algorithmes stochastiques, Mathématiques et Applications, Springer Berlin Heidelberg, 1996.
[14] V. Fabian, On asymptotic normality in stochastic approximation, The Annals of Mathematical Statistics, (1968), pp. 1327–1332.
[15] O. Fercoq and P. Richtárik, Optimization in high dimensions via accelerated, parallel, and proximal coordinate descent, Siam review, 58 (2016), pp. 739–771.
[16] R. Gower, N. Le Roux, and F. Bach, Tracking the gradients using the hessian: A new look at variance reducing stochastic methods, in International Conference on Artificial Intelligence and Statistics, PMLR, 2018, pp. 707–715.
[17] R. M. Gower, M. Schmidt, F. Bach, and P. Richtárik, Variance-reduced methods for machine learning, Proceedings of the IEEE, 108 (2020), pp. 1968–1983.
[18] C. Guille-Escuret, A. Ibrahim, B. Goujaud, and I. Mitliagkas, Gradient descent is optimal under lower restricted secant inequality and upper error bound, Advances in Neural Information Processing Systems, 35 (2022), pp. 24893–24904.
[19] H. Karimi, J. Nutini, and M. Schmidt, Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition, in Joint European conference on machine learning and knowledge discovery in databases, Springer, 2016, pp. 795–811.
[20] J. Kiefer and J. Wolfowitz, Stochastic estimation of the maximum of a regression function, The Annals of Mathematical Statistics, (1952), pp. 462–466.
[21] H. J. Kushner and H. Huang, Rates of convergence for stochastic approximation type algorithms, SIAM Journal on Control and Optimization, 17 (1979), pp. 607–617.
[22] H. J. Kushner and G. Yin, Stochastic Approximation and Recursive Algorithms and Applications, Springer New York, NY, 2003.
[23] C. K. Lauand, I. Kontoyiannis, and S. Meyn, The case for and against fixed step-size: Stochastic approximation algorithms in optimization and machine learning, 2025, https://arxiv.org/abs/2309.02944, https://arxiv.org/abs/2309.02944.
[24] R. Leluc and F. Portier, Sgd with coordinate sampling: Theory and practice, Journal of Machine Learning Research, 23 (2022), pp. 1–47.
[25] J. Liu and Y. Yuan, On almost sure convergence rates of stochastic gradient methods, in Conference on Learning Theory, PMLR, 2022, pp. 2963–2983.
[26] L. Nguyen, P. H. Nguyen, M. Dijk, P. Richtárik, K. Scheinberg, and M. Takác, Sgd and hogwild! convergence without the bounded gradients assumption, in International Conference on Machine Learning, PMLR, 2018, pp. 3750–3758.
[27] B. Palaniappan and F. Bach, Stochastic variance reduction methods for saddle-point problems, Advances in Neural Information Processing Systems, 29 (2016).
[28] M. Pelletier, On the almost sure asymptotic behaviour of stochastic algorithms, Stochastic processes and their applications, 78 (1998), pp. 217–244.
[29] M. Pelletier, Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing, Annals of Applied Probability, (1998), pp. 10–44.
[30] B. T. Polyak and A. B. Juditsky, Acceleration of stochastic approximation by averaging, SIAM journal on control and optimization, 30 (1992), pp. 838–855.
[31] C. Poon, J. Liang, and C. Schoenlieb, Local convergence properties of saga/prox-svrg and acceleration, in International Conference on Machine Learning, PMLR, 2018, pp. 4124–4132.
[32] X. Qian, Z. Qu, and P. Richtárik, Saga with arbitrary sampling, in International Conference on Machine Learning, PMLR, 2019, pp. 5190–5199.
[33] H. Robbins and S. Monro, A stochastic approximation method, The annals of mathematical statistics, (1951), pp. 400–407.
[34] H. Robbins and D. Siegmund, A convergence theorem for non negative almost supermartingales and some applications, in Optimizing methods in statistics, Elsevier, 1971, pp. 233–257.
[35] N. Roux, M. Schmidt, and F. Bach, A stochastic gradient method with an exponential convergence _rate for finite training sets, Advances in neural information processing systems, 25 (2012).
[36] J. Sacks, Asymptotic distribution of stochastic approximation procedures, The Annals of Mathematical Statistics, 29 (1958), pp. 373–405.
[37] M. Schmidt, N. Le Roux, and F. Bach, Minimizing finite sums with the stochastic average gradient, Mathematical Programming, 162 (2017), pp. 83–112.
[38] L. Xiao and T. Zhang, A proximal stochastic gradient method with progressive variance

reduction, SIAM Journal on Optimization, 24 (2014), pp. 2057–2075.
[39] X. Yi, S. Zhang, T. Yang, T. Chai, and K. H. Johansson, Exponential convergence for distributed optimization under the restricted secant inequality condition, IFAC-PapersOnLine, 53 (2020), pp. 2672–2677.
[40] L. Zhang and L. Guo, Asymptotically efficient adaptive identification under saturated output observations, SIAM Journal on Control and Optimization, 63 (2025), pp. 2338–2368.
[41] L.-X. Zhang, Central limit theorems of a recursive stochastic algorithm with applications to adaptive designs, The Annals of Applied Probability, 26 (2016), pp. 3630–3658.