Submitted to the Annals of Statistics

BRIDGING THE GAP BETWEEN CONSTANT STEP SIZE STOCHASTIC GRADIENT DESCENT AND MARKOV CHAINS

BY AYMERIC DIEULEVEUT *,† ALAIN DURMUS ‡ AND FRANCIS BACH *

* *INRIA - Département d'informatique de l'ENS, École normale supérieure, CNRS, PSL Research University*
† *EPFL - Machine Learning and Optimization lab*
‡ *CMLA - École normale supérieure Paris-Saclay, CNRS, Université Paris-Saclay*

Abstract: We consider the minimization of a strongly convex objective function given access to unbiased estimates of its gradient through stochastic gradient descent (SGD) with constant step-size. While the detailed analysis was only performed for quadratic functions, we provide an explicit asymptotic expansion of the moments of the averaged SGD iterates that outlines the dependence on initial conditions, the effect of noise and the step-size, as well as the lack of convergence in the general (non-quadratic) case. For this analysis, we bring tools from Markov chain theory into the analysis of stochastic gradient. We then show that Richardson-Romberg extrapolation may be used to get closer to the global optimum and we show empirical improvements of the new extrapolation scheme.

1. Introduction. We consider the minimization of an objective function given access to unbiased estimates of the function gradients. This key methodological problem has raised interest in different communities: in large-scale machine learning [9, 51, 52], optimization [42, 44], and stochastic approximation [28, 46, 50]. The most widely used algorithms are stochastic gradient descent (SGD), a.k.a. Robbins-Monro algorithm [49], and some of its modifications based on averaging of the iterates [46, 48, 53].

While the choice of the step-size may be done robustly in the deterministic case (see *e.g.* [8]), this remains a traditional theoretical and practical issue in the stochastic case. Indeed, early work suggested to use step-sizes decaying with the number k of iterations as $O(1/k)$ [49], but it appeared to be non-robust to ill-conditioning and slower decays such as $O(1/\sqrt{k})$ together with averaging lead to both good practical and theoretical performance [3, 42].

AMS 2000 subject classifications: primary 62L20; secondary 90C15, 93E35
Keywords and phrases: stochastic gradient descent, Markov chains

We consider in this paper constant step-size SGD, which is often used in practice. Although the algorithm is not converging in general to the global optimum of the objective function, constant step-sizes come with benefits: (a) there is a single parameter value to set as opposed to the several choices of parameters to deal with decaying step-sizes, *e.g.* as $1/(\square k + \triangle)^{\circ}$; the initial conditions are forgotten exponentially fast[1] for well-conditioned (*e.g.* strongly convex) problems [40, 41], and the performance, although not optimal, is sufficient in practice (in a machine learning set-up, being only 0.1% away from the optimal prediction often does not matter).

The main goals of this paper are (a) to gain a complete understanding of the properties of constant-step-size SGD in the strongly convex case, and (b) to propose provable improvements to get closer to the optimum when precision matters or in high-dimensional settings. We consider the iterates of the SGD recursion on $\mathbb{R}^d$ defined starting from $\theta_0 \in \mathbb{R}^d$, for $k \geq 0$, and a step-size $\gamma > 0$ by

$$\theta_{k+1}^{(\gamma)} = \theta_k^{(\gamma)} - \gamma\big[f'(\theta_k^{(\gamma)}) + \varepsilon_{k+1}(\theta_k^{(\gamma)})\big] \,, \tag{1}$$

where f is the objective function to minimize (in machine learning the generalization performance), $\varepsilon_{k+1}(\theta_k^{(\gamma)})$ the zero-mean statistically independent noise (in machine learning, obtained from a single observation). Following [5], we leverage the property that the sequence of iterates $(\theta_k^{(\gamma)})_{k\geq 0}$ is a *homogeneous Markov chain*.

This interpretation allows us to capture the general behavior of the algorithm. In the strongly convex case, this Markov chain converges exponentially fast to a unique stationary distribution π_γ (see Proposition 2) highlighting the facts that (a) initial conditions of the algorithms are forgotten quickly, and (b) the algorithm does not converge to a point but oscillates around the mean of π_γ. See an illustration in Figure 1 (left). It is known that the oscillations of the non-averaged iterates have an average magnitude of $\gamma^{1/2}$ [45].

Consider the process $(\bar{\theta}_k^{(\gamma)})_{k\geq 0}$ given for all $k \geq 0$ by

$$\bar{\theta}_k^{(\gamma)} = \frac{1}{k+1}\sum_{j=0}^{k} \theta_j^{(\gamma)} \,. \tag{2}$$

Then under appropriate conditions on the Markov chain $(\theta_k^{(\gamma)})_{k\geq 0}$, a central limit theorem on $(\bar{\theta}_k^{(\gamma)})_{k\geq 0}$ holds which implies that $\bar{\theta}_k^{(\gamma)}$ converges at rate

[1] On the contrary, step size scaling as $1/(\mu k)$ (with μ the strong convexity constant) forget the initial condition much slower. They also require to access μ (which may be difficult) and are very sensitive to its mis-specification [52].

$O(1/\sqrt{k})$ to

$$\bar{\theta}_\gamma = \int_{\mathbb{R}^d} \vartheta \, \mathrm{d}\pi_\gamma(\vartheta) \,. \tag{3}$$

The deviation between $\bar{\theta}_k^{(\gamma)}$ and the global optimum θ^* is thus composed of a stochastic part $\bar{\theta}_k^{(\gamma)} - \bar{\theta}_\gamma$ and a deterministic part $\bar{\theta}_\gamma - \theta^*$.

For quadratic functions, it turns out that the deterministic part vanishes [5], that is, $\bar{\theta}_\gamma = \theta^*$ and thus averaged SGD with a constant step-size does converge. However, it is not true for general objective functions where we can only show that $\bar{\theta}_\gamma - \theta^* = O(\gamma)$, and this deviation is the reason why constant step-size SGD is not convergent.

The first main contribution of the paper is to provide an explicit asymptotic expansion in the step-size γ of $\bar{\theta}_\gamma - \theta^*$. Second, a quantitative version of a central limit theorem is established which gives a bound on $\mathbb{E}[\|\bar{\theta}_\gamma - \bar{\theta}_k^{(\gamma)}\|^2]$ that highlights all dependencies on initial conditions and noise variance, as achieved for least-squares by [15], with an explicit decomposition into "bias" and "variance" terms: the bias term characterizes how fast initial conditions are forgotten and is proportional to $\mathrm{N}(\theta_0 - \theta^*)$, for a suitable norm $\mathrm{N} : \mathbb{R}^d \to \mathbb{R}_+$; while the variance term characterizes the effect of the noise in the gradient, independently of the starting point, and increases with the covariance of the noise.

Moreover, akin to weak error results for ergodic diffusions [57], we achieve a non-asymptotic weak error expansion in the step-size between π_γ and the Dirac measure on $\mathbb{R}^d$ concentrated at θ^*. Namely, we prove that for all functions $g : \mathbb{R}^d \to \mathbb{R}$, regular enough, $\int_{\mathbb{R}^d} g(\vartheta)\mathrm{d}\pi_\gamma(\vartheta) = g(\theta^*) + \gamma C_1^g + r_\gamma^g$, $r_\gamma^g \in \mathbb{R}^d$, $\|r_\gamma^g\| \leq C_2^g \gamma^2$, for some $C_1^g, C_2^g \geq 0$ independent of γ. Given this expansion, we can now use a very simple trick from numerical analysis, namely Richardson-Romberg extrapolation [54]: if we run two SGD recursions $(\theta_k^{(\gamma)})_{k\geq 0}$ and $(\theta_k^{(2\gamma)})_{k\geq 0}$ with the two different step-sizes γ and 2γ, then the average processes $(\bar{\theta}_k^{(\gamma)})_{k\geq 0}$ and $(\bar{\theta}_k^{(2\gamma)})_{k\geq 0}$ will converge to $\bar{\theta}_\gamma$ and $\bar{\theta}_{2\gamma}$ respectively. Since $\bar{\theta}_\gamma = \theta^* + \gamma\Delta_1^{\mathrm{Id}} + r_\gamma^{\mathrm{Id}}$ and $\bar{\theta}_{2\gamma} = \theta^* + 2\gamma\Delta_1^{\mathrm{Id}} + r_{2\gamma}^{\mathrm{Id}}$, for $r_\gamma^{\mathrm{Id}}, r_{2\gamma}^{\mathrm{Id}} \in \mathbb{R}^d$, $\max(\|2r_\gamma^{\mathrm{Id}}\|, \|r_{2\gamma}^{\mathrm{Id}}\|) \leq 2C\gamma^2$, for $C \geq 0$ and $\Delta \in \mathbb{R}^d$ independent of γ, the combined iterates $2\bar{\theta}_k^{(\gamma)} - \bar{\theta}_k^{(2\gamma)}$ will converge to $\theta^* + 2r_\gamma^{\mathrm{Id}} - r_{2\gamma}^{\mathrm{Id}}$ which is closer to θ^* by a factor γ. See illustration in Figure 1(right).

In summary, we make the following contributions:

- We provide in Section 2 an asymptotic expansion in γ of $\bar{\theta}_\gamma - \theta^*$ and an explicit version of a central limit theorem is given which bounds $\mathbb{E}[\|\bar{\theta}_\gamma - \bar{\theta}_k^{(\gamma)}\|^2]$. These two results outline the dependence on initial conditions, the effect of noise and the step-size.

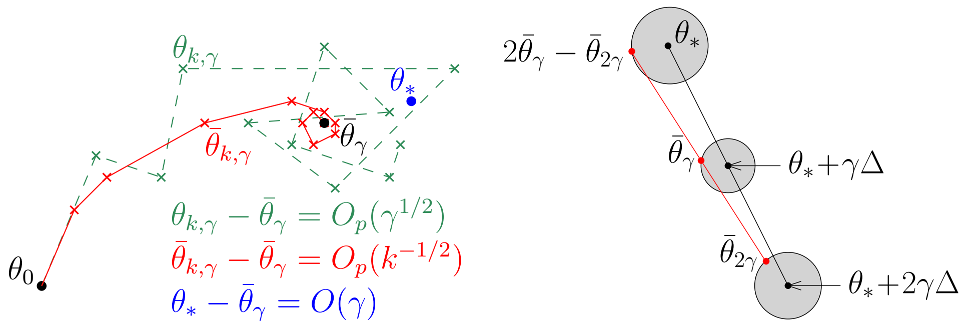

Figure 1: (Left) Convergence of iterates $\theta_k^{(\gamma)}$ and averaged iterates $\bar{\theta}_k^{(\gamma)}$ to the mean $\bar{\theta}_\gamma$ under the stationary distribution π_γ. (Right) Richardson-Romberg extrapolation, the disks are of radius $O(\gamma^2)$.

- We show in Section 2 that Richardson-Romberg extrapolation may be used to get closer to the global optimum.
- We borrow and adapt in Section 3 some techniques to analyze asymptotic bias of numerical schemes in the context of diffusion processes to get new insight about SGD. We believe that this analogy and the associated ideas are interesting in their own right.
- We show in Section 4 empirical improvements of the extrapolation schemes.

These results can be used directly in practice, to achieve faster convergence in both asymptotic and non asymptotic regimes. Moreover, convergence results can be used to derive confidence intervals for θ^*, as in [13, 55]. Another important application is the design of automatic restart schemes for SGD: in applications (especially in non-convex settings), practitioners typically use epoch-wise constant step size: the step size is periodically reduced [**?**, **?**]. However, the reduction scheduling is typically hand-tuned, which is a major burden. Automatic restart strategies have been considered [11], they are based on reducing the step size when stationarity is reached. The detailed analysis of stationarity we provide can allow to design new or more efficient restart strategies for such applications.

Notations. We first introduce several notations. We consider the finite dimensional[2] Euclidean space $\mathbb{R}^d$ embedded with its canonical inner product $\langle \cdot, \cdot \rangle$. Denote by $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$ the canonical basis of $\mathbb{R}^d$. Let E and F be two real vector spaces, we denote by $E \otimes F$ the tensor product of E and F. For

[2] Proofs and results could be extended to an infinite dimensional domain. However, it would require heavy technical considerations without bringing new important insights.

all $x \in E$ and $y \in F$ denote by $x \otimes y \in E \otimes F$ the tensor product of x and y. Denote by $E^{\otimes k}$ the k^{th} tensor power of E and $x^{\otimes k} \in E^{\otimes k}$ the k^{th} tensor power of x. We let $\mathcal{L}((\mathbb{R}^d)^{\otimes k}, \mathbb{R}^\ell)$ stand for the set of linear maps from $(\mathbb{R}^n)^{\otimes k}$ to $\mathbb{R}^\ell$ and for $\mathrm{L} \in \mathcal{L}((\mathbb{R}^d)^{\otimes k}, \mathbb{R}^\ell)$, we denote by $\|\mathrm{L}\|$ the operator norm of L.

Let $n \in \mathbb{N}^*$, denote by $\mathrm{C}^n(\mathbb{R}^d, \mathbb{R}^m)$ the set of n times continuously differentiable functions from $\mathbb{R}^d$ to $\mathbb{R}^m$. Let $F \in \mathrm{C}^n(\mathbb{R}^d, \mathbb{R}^m)$, denote by $F^{(n)}$ or $D^n F$, the n^{th} differential of f. Let $f \in \mathrm{C}^n(\mathbb{R}^d, \mathbb{R})$. For any $x \in \mathbb{R}^d$, $f^{(n)}(x)$ is a tensor of order n. For example, for all $x \in \mathbb{R}^d$, $f^{(3)}(x)$ is a third order tensor. In addition, for any $x \in \mathbb{R}^d$ and any matrix, $M \in \mathbb{R}^{d \times d}$, we define $f^{(3)}(x)M$ as the vector in $\mathbb{R}^d$ given by: for any $l \in \{1, \dots, d\}$, the l^{th} coordinate is given by $(f^{(3)}(x)M)_l = \sum_{i,j=1}^d M_{i,j} \frac{\partial^3 f}{\partial x_i \partial x_j \partial x_l}(x)$. By abuse of notations, for $f \in \mathrm{C}^1(\mathbb{R}^d)$, we identify f' with the gradient of f and if $f \in \mathrm{C}^2(\mathbb{R}^d)$, we identify f'' with the Hessian matrix of f. A function $f : \mathbb{R}^d \to \mathbb{R}^q$ is said to be locally Lipschitz with polynomial growth or pseudo-Lipschitz if there exists $\alpha \geq 0$ and $C \geq 0$ such that for all $x, y \in \mathbb{R}^d$, $\|f(x) - f(y)\| \leq C(1 + \|x\|^\alpha + \|y\|^\alpha) \|x - y\|$. In this document, any locally Lipschitz function is assumed to be locally Lipschitz with polynomial growth and therefore for ease of presentation, we do not specify it in the sequel. For ease of notations and depending on the context, we consider $M \in \mathbb{R}^{d \times d}$ either as a matrix or a second order tensor. More generally, any $M \in \mathrm{L}((\mathbb{R}^d)^{\otimes k}, \mathbb{R})$ will be also consider as an element of $\mathrm{L}((\mathbb{R}^d)^{\otimes (k-1)}, \mathbb{R}^d)$ by the canonical bijection. Besides, For any matrices $M, N \in \mathbb{R}^{d \times d}$, $M \otimes N$ is defined as the endomorphism of $\mathbb{R}^{d \times d}$ such that $M \otimes N : P \mapsto MPN$. For any matrix $M \in \mathbb{R}^{d \times d}$, $\mathrm{tr}(M)$ is the trace of M, *i.e.* the sum of diagonal elements of the matrix M.

For $a, b \in \mathbb{R}$, denote by $a \vee b$ and $a \wedge b$ the maximum and the minimum of a and b respectively. Denote by $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ the floor and ceiling function respectively.

Denote by $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-field of $\mathbb{R}^d$. For all $x \in \mathbb{R}^d$, $\updelta_x$ stands for the Dirac measure at x.

2. Main results.

In this section, we describe the assumptions underlying our analysis, describe our main results and their implications.

2.1. *Setting.*

Let $f : \mathbb{R}^d \to \mathbb{R}$ be an objective function, satisfying the following assumptions:

A1. *The function f is strongly convex with convexity constant $\mu > 0$,* *i.e.* *for all* $\theta_1, \theta_2 \in \mathbb{R}^d$ *and* $t \in [0,1]$,

$$f(t\theta_1 + (1-t)\theta_2) \le tf(\theta_1) + (1-t)f(\theta_2) - (\mu/2)t(1-t)\|\theta_1 - \theta_2\|^2 \ .$$

A2. *The function f is five times continuously differentiable with second to fifth uniformly bounded derivatives: for all $k \in \{2, \ldots, 5\}$, $\sup_{\theta \in \mathbb{R}^d} \left\| f^{(k)}(\theta) \right\| < +\infty$. Especially f is L-smooth with $L \ge 0$: for all $\theta_1, \theta_2 \in \mathbb{R}^d$*

$$\left\| f'(\theta_1) - f'(\theta_2) \right\| \le L \|\theta_1 - \theta_2\| \ .$$

If there exists a positive definite matrix $\Sigma \in \mathbb{R}^{d\times d}$, such that the function f is the quadratic function $\theta \mapsto \|\Sigma^{1/2}(\theta - \theta^*)\|^2/2$, then **A**1, **A**2 are satisfied.

In the definition of SGD given by (1), $(\varepsilon_k)_{k\ge1}$ is a sequence of random functions from $\mathbb{R}^d$ to $\mathbb{R}^d$ satisfying the following properties.

A3. *There exists a filtration $(\mathcal{F}_k)_{k\ge0}$ (i.e. for all $k \in \mathbb{N}$, $\mathcal{F}_k \subset \mathcal{F}_{k+1}$) on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that for any $k \in \mathbb{N}$ and $\theta \in \mathbb{R}^d$, $\varepsilon_{k+1}(\theta)$ is a $\mathcal{F}_{k+1}$-measurable random variable and $\mathbb{E}\left[\varepsilon_{k+1}(\theta)|\mathcal{F}_k\right] = 0$. In addition, $(\varepsilon_k)_{k\in\mathbb{N}^*}$ are independent and identically distributed (i.i.d.) random fields. Moreover, we assume that θ_0 is $\mathcal{F}_0$-measurable.*

A3 expresses that we have access to an i.i.d. sequence $(f'_k)_{k\in\mathbb{N}^*}$ of unbiased estimator of f', *i.e.* for all $k \in \mathbb{N}$ and $\theta \in \mathbb{R}^d$,

$$f'_{k+1}(\theta) = f'(\theta) + \varepsilon_{k+1}(\theta) \ . \tag{4}$$

Note that we do not assume random vectors $(\varepsilon_{k+1}(\theta_k^{(\gamma)}))_{k\in\mathbb{N}}$ to be i.i.d., a stronger assumption generally referred to as the semi-stochastic. Moreover, as θ_0 is $\mathcal{F}_0$-measurable, for any $k \in \mathbb{N}$, θ_k is $\mathcal{F}_k$-measurable.

We also consider the following conditions on the noise, for $p \ge 2$:

A4 (p). *For any $k \in \mathbb{N}^*$, f'_k is almost surely L-co-coercive (with the same constant as in **A**2): that is, for any $\eta, \theta \in \mathbb{R}^d$, $L\langle f'_k(\theta) - f'_k(\eta), \theta - \eta\rangle \ge \|f'_k(\theta) - f'_k(\eta)\|^2$. Moreover, there exists $\tau_p \ge 0$, such that for any $k \in \mathbb{N}^*$, $\mathbb{E}^{1/p}[\|\varepsilon_k(\theta^*)\|^p] \le \tau_p$.*

Almost sure L-co-coercivity [60] is for example satisfied if for any $k \in \mathbb{N}^*$, there exists a random function f_k such that $f'_k = (f_k)'$ and which is a.s. convex and L-smooth. Weaker assumptions on the noise are discussed in Section 6.1. Finally we emphasize that under **A**3, in order to verify that **A**4(p) holds, $p \ge 2$, it suffices to show that f'_1 is almost surely L-co-coercive and $\mathbb{E}^{1/p}[\|\varepsilon_1(\theta^*)\|^p] \le \tau_p$. Under **A**3-**A**4(2), consider the function $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ defined for all $\theta \in \mathbb{R}^d$ by

$$\mathcal{C}(\theta) = \mathbb{E}\left[\varepsilon_1(\theta)^{\otimes 2}\right] \ . \tag{5}$$

A5. *The function $\mathcal{C}$ is three time continuously differentiable and there exist $M_\varepsilon, k_\varepsilon \geq 0$ such that for all $\theta \in \mathbb{R}^d$,*

$$\max_{i \in \{1,2,3\}} \left\| \mathcal{C}^{(i)}(\theta) \right\| \leq M_\varepsilon \left\{ 1 + \|\theta - \theta^*\|^{k_\varepsilon} \right\} .$$

In other words, we assume that the covariance matrix $\theta \mapsto C(\theta)$ is a regular enough function, which is satisfied in natural settings.

EXAMPLE 1 (Learning from i.i.d. observations). *Our main motivation comes from machine learning; consider two sets $\mathcal{X}, \mathcal{Y}$ and a convex loss function $\mathrm{L} : \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}$. The objective function is the generalization error $f_{\mathrm{L}}(\theta) = \mathbb{E}_{X,Y}[\mathrm{L}(X, Y, \theta)]$, where (X, Y) are some random variables. Given i.i.d. observations $(X_k, Y_k)_{k \in \mathbb{N}^*}$ with the same distribution as (X, Y), for any $k \in \mathbb{N}^*$, we define $f_k(\cdot) = \mathrm{L}(X_k, Y_k, \cdot)$ the loss with respect to observation k. SGD then corresponds to following gradient of the loss on a single independent observation (X_k, Y_k) at each step; Assumption **A**3 is then satisfied with $\mathcal{F}_k = \sigma((X_j, Y_j)_{j \in \{1,\dots,k\}})$.*

Two classical situations are worth mentioning. On the first hand, in least-squares regression, *$\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, and the loss function is* $\mathrm{L}(X, Y, \theta) = (\langle X, \theta \rangle - Y)^2$. *Then f_Σ is the quadratic function $\theta \mapsto \|\Sigma^{1/2}(\theta - \theta^*)\|^2/2$, with $\Sigma = \mathbb{E}[XX^\top]$, which satisfies Assumption **A**2. For any $\theta \in \mathbb{R}^d$,*

$$\varepsilon_k(\theta) = X_k X_k^\top \theta - X_k Y_k \ . \tag{6}$$

*Then, for any $p \geq 2$, Assumption **A**4(p) and **A**5 is satisfied as soon as the observations are a.s. bounded, while **A**1 is satisfied if the second moment matrix is invertible or additional regularization is added. In this setting, ε_k can be decomposed as $\varepsilon_k = \varrho_k + \xi_k$ where ϱ_k is the multiplicative part, ξ_k the additive part, given for $\theta \in \mathbb{R}^d$ by $\varrho_k(\theta) = (X_k X_k^\top - \Sigma)(\theta - \theta^*)$ and*

$$\xi_k = (X_k^\top \theta^* - Y_k) X_k \ . \tag{7}$$

For all $k \geq 1$, ξ_k does not depend on θ. These two parts in the noise will appear in Corollary 6. Finally assume that there exists $r \geq 0$ such that

$$\mathbb{E}[\|X_k\|^2 \, X_k X_k^\top] \preccurlyeq r^2 \Sigma \ , \tag{8}$$

*then **A**4(4) is satisfied. This assumption is satisfied, e.g., for a.s. bounded data, or for data with bounded kurtosis, see [18] for details.*

On the other hand, in (regularized) logistic regression, *where* $\mathrm{L}(X, Y, \theta) = \log(1 + \exp(-Y \langle X, \theta \rangle))$, *Assumptions **A**4 or **A**2 are similarly satisfied, while **A**1 holds when regularization is added, or with an additional restriction to a compact set (using self-concordance assumptions [3] would allow a direct unconstrained application).*

2.2. *Summary and discussion of main results.* Under the stated assumptions, for all $\gamma \in (0, 2/L)$ and $\theta_0 \in \mathbb{R}^d$, the Markov chain $(\theta_k^{(\gamma)})_{k\geq 0}$ converges in a certain sense specified below to a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, π_γ satisfying $\int_{\mathbb{R}^d} \|\vartheta\|^2 \pi_\gamma(\mathrm{d}\vartheta) < +\infty$, see Proposition 2 in Section 3. In the next section, by two different methods (Theorem 4 and Theorem 7), we show that under suitable conditions on f and the noise $(\varepsilon_k)_{k\geq 1}$, there exists $\Delta \in \mathbb{R}^d$ such that for all small enough $\gamma \geq 0$,

$$\bar{\theta}_\gamma = \int_{\mathbb{R}^d} \vartheta \pi_\gamma(\mathrm{d}\vartheta) = \theta^* + \gamma\Delta + r_\gamma^{(1)} ,$$

where $r_\gamma^{(1)} \in \mathbb{R}^d$, $\|r_\gamma^{(1)}\| \leq C\gamma^2$ for some constant $C \geq 0$ independent of γ. Using Proposition 2, we get that for all $k \geq 1$,

$$\mathbb{E}[\bar{\theta}_k^{(\gamma)} - \theta^*] = \frac{A(\theta_0, \gamma)}{k} + \gamma\Delta + r_\gamma^{(2)} , \tag{9}$$

where $r_\gamma^{(2)} \in \mathbb{R}^d$, $\|r_\gamma^{(2)}\| \leq C(\gamma^2 + \mathrm{e}^{-k\mu\gamma})$ for some constant $C \geq 0$ independent of γ.

This expansion in the step-size γ shows that a Richardson-Romberg extrapolation can be used to have better estimates of θ^*. Consider the average iterates $(\bar{\theta}_{2\gamma}^{(k)})_{k\geq 0}$ and $(\bar{\theta}_k^{(\gamma)})_{k\geq 0}$ associated with SGD with step size 2γ and γ respectively. Then (9) shows that $(2\bar{\theta}_k^{(\gamma)} - \bar{\theta}_k^{(2\gamma)})_{k\geq 0}$ satisfies

$$\mathbb{E}[2\bar{\theta}_k^{(\gamma)} - \bar{\theta}_k^{(2\gamma)} - \theta^*] = \frac{2A(\theta_0, \gamma) - A(\theta_0, 2\gamma)}{k} + 2r_\gamma^{(2)} - r_{2\gamma}^{(2)} ,$$

and therefore is closer to the optimum θ^*. This very simple trick improves the convergence by a factor of γ (at the expense of a slight increase of the variance). In practice, while the objective values at the un-averaged gradient iterates $\theta_k^{(\gamma)}$ saturate (*i.e.* stop decaying) at a suboptimal value rapidly, $\bar{\theta}_k^{(\gamma)}$ may already perform well enough to avoid saturation on real data-sets [5]. The Richardson-Romberg extrapolated iterate $2\bar{\theta}_k^{(\gamma)} - \bar{\theta}_k^{(2\gamma)}$ very rarely reaches saturation in practice. This appears in synthetic experiments presented in Section 4. Moreover, this procedure only requires to compute two parallel SGD recursions, either with the same inputs, or with different ones, and is naturally parallelizable.

In Section 3.2, we give a quantitative version of a central limit theorem for $(\bar{\theta}_k^{(\gamma)})_{k\geq 0}$, for a fixed $\gamma > 0$ and k going to $+\infty$: under appropriate conditions, there exist constants $B_1(\gamma)$ and $B_2(\gamma)$ such that

$$\mathbb{E}\left[\left\|\bar{\theta}_k^{(\gamma)} - \bar{\theta}_\gamma\right\|^2\right] = B_1(\gamma)/k + B_2(\gamma)/k^2 + O(1/k^3) . \tag{10}$$

Combining (9) and (10) characterizes the bias/variance trade-off of SGD used to estimate θ^*.

2.3. *Related work.* The idea to study stochastic approximation algorithms using results and techniques from the Markov chain literature is not new. It goes back to [23], which shows under appropriate conditions that solutions of stochastic differential equations (SDE)

$$\mathrm{d}Y_t = -f'(Y_t)\mathrm{d}t + \gamma_t \mathrm{d}B_t \ ,$$

where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion and $(\gamma_t)_{t\geq 0}$ is a one-dimensional positive function, $\lim_{t\to+\infty} \gamma_t = 0$, converge in probability to some minima of f. Another example is [47] which extends the classical Foster-Lyapunov criterion from Markov chain theory (see [38]) to study the stability of the Least Mean Square algorithm. In [10], the authors are interested in the convergence of the multidimensional Kohonen algorithm. They show that the Markov chain defined by this algorithm is uniformly ergodic and derive asymptotic properties on its limiting distribution.

The techniques we use in this paper to establish our results share a lot of similarities with previous work. For example, our first results in Section 3.1 and Section 3.2 can be seen as complementary results of [2]. Indeed, in [2], the authors decompose the tracking error of a general algorithm in a linear regression model. To prove their result, they develop the error using a perturbation approach. However, for linear regression, $\bar{\theta}_\gamma = \theta^*$, which justifies the present work which deals with potentially non-quadratic objective functions f.

Another and significant point of view to study stochastic approximation relies on the gradient flow equation associated with the vector field f': $\dot{x}_t = -f'(x_t)$. This approach was introduced by [31] and [28] and has been applied in numerous papers since then, see [36, 37, 7, 6, 56]. To establish our results in Section 3.3, we use the strong connection between SGD and the gradient flow equation as well, in particular we introduce the Poisson solution associated with the gradient flow equation. The combination of the relation between stochastic approximation algorithms with the gradient flow equation and the Markov chain theory has been developed in [21] and [22]. In particular, [22] establishes under appropriate conditions that there exists for all $\gamma \in (0, \gamma_0)$, with γ_0 small enough, an invariant distribution π_γ for the Markov chain $(\theta_k^{(\gamma)})_{k\in\mathbb{N}}$, and $(\pi_\gamma)_{\gamma\in(0,\gamma_0)}$ is tight. In addition, they show that any limiting distributions is invariant for the gradient flow associated with f'. Note that their conditions and results are different from ours. In particular, we do not assume that $(\theta_k^{(\gamma)})_{k\in\mathbb{N}}$ is Feller but require that f is strongly convex contrary to [22]. In addition, we establish an explicit expansion in the step size γ for $\bar{\theta}_\gamma - \theta^*$ and more generally for the weak error between π_γ and δ_{θ^*}.

To the authors knowledge, the use of the Richardson-Romberg method for stochastic approximation has only been considered in [39] to recover the minimax rate for recursive estimation of time varying autoregressive process.

Several attempts have been made to improve convergence of SGD. [5] proposed an online Newton algorithm which converges in practice to the optimal point with constant step-size but has no convergence guarantees. The quadratic case was studied by [5], for the (uniform) average iterate: the variance term is upper bounded by $\sigma^2 d/n$ and the squared bias term by $\|\theta^*\|^2/(\gamma n)$. This last term was improved to $\|\Sigma^{-1/2}\theta^*\|^2/(\gamma n)^2$ by [15, 16], showing that asymptotically, the bias term is negligible, see also [29]. Analysis has been extended to "tail averaging" [26], to improve the dependence on the initial conditions. Note that this procedure can be seen as a Richardson-Romberg trick with respect to k. Other strategies were suggested to improve the speed at which initial conditions were forgotten, for example using acceleration when the noise is additive [18, 27]. A criterion to check when SGD with constant step size is close to its limit distribution was recently proposed in [11].

In the context of discretization of ergodic diffusions, weak error estimates between the stationary distribution of the discretization and the invariant distribution of the associated diffusion have been first shown by [57] and [35] in the case of the Euler-Maruyama scheme. Then, [57] suggested the use of Richardson-Romberg interpolation to improve the accuracy of estimates of integrals with respect to the invariant distribution of the diffusion. Extension of these results have been obtained for other types of discretization by [1] and [12]. We show in Section 3.3 that a weak error expansion in the step-size γ also holds for SGD between π_γ and δ_{θ^*}. Interestingly as to the Euler-Maruyama discretization, SGD has a weak error of order γ. In addition, [19] proposed and analyzed the use of Richardson-Romberg extrapolation applied to the stochastic gradient Langevin dynamics (SGLD) algorithm. This method introduced by [59] combines SGD and the Euler-Maruyama discretization of the Langevin diffusion associated to a target probability measure [14, 20]. Note that this method is however completely different from SGD, in part because Gaussian noise of order $\gamma^{1/2}$ (instead of γ) is injected in SGD which changes the overall dynamics.

Finally, it is worth mentioning [33, 34] which are interested in showing that the invariant measure of constant step-size SGD for an appropriate choice of the step-size γ, can be used as a proxy to approximate the target distribution π with density with respect to the Lebesgue measure e^{-f}. Note that the perspective and purpose of this paper is completely different since we are interested in optimizing the function f and not in sampling from π.

3. Detailed analysis. In this Section, we describe in detail our approach. A first step is to describe the existence of a unique stationary distribution π_γ for the Markov chain $(\theta_k^{(\gamma)})_{k\geq 0}$ and the convergence of this Markov chain to π_γ in the Wasserstein distance of order 2.

Limit distribution. We cast in this section SGD in the Markov chain framework and introduce basic notion related to this theory, see [38] for an introduction to this topic. Consider the Markov kernel R_γ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ associated with SGD iterates $(\theta_k^{(\gamma)})_{k\in\mathbb{N}}$, *i.e.* for all $k \in \mathbb{N}$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, almost surely $R_\gamma(\theta_k, \mathsf{A}) = \mathbb{P}(\theta_{k+1} \in \mathsf{A}|\theta_k)$, for all $\theta_0 \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $\theta \mapsto R_\gamma(\theta, \mathsf{A})$ is Borel measurable and $R_\gamma(\theta_0, \cdot)$ is a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. For all $k \in \mathbb{N}^*$, we define the Markov kernel R_γ^k recursively by $R_\gamma^1 = R_\gamma$ and for $k \geq 1$, for all $\theta_0 \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$

$$R_\gamma^{k+1}(\theta_0, \mathsf{A}) = \int_{\mathbb{R}^d} R_\gamma^k(\theta_0, \mathrm{d}\theta) R_\gamma(\theta, \mathsf{A}) \,.$$

For any probability measure λ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, we define the probability measure λR_γ for all $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$\lambda R_\gamma^k(\mathsf{A}) = \int_{\mathbb{R}^d} \lambda(\mathrm{d}\theta) R_\gamma^k(\theta, \mathsf{A}) \,.$$

By definition, for any probability measure λ on $\mathcal{B}(\mathbb{R}^d)$ and $k \in \mathbb{N}^*$, λR_γ^k is the distribution of $\theta_k^{(\gamma)}$ started from θ_0 drawn from λ. For any function $\phi : \mathbb{R}^d \to \mathbb{R}_+$ and $k \in \mathbb{N}^*$, define the measurable function $R_\gamma^k \phi : \mathbb{R}^d \to \mathbb{R}$ for all $\theta_0 \in \mathbb{R}^d$,

$$R_\gamma^k \phi(\theta_0) = \int_{\mathbb{R}^d} \phi(\theta) R_\gamma^k(\theta_0, \mathrm{d}\theta) \,.$$

For any measure λ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and any measurable function $h : \mathbb{R}^d \to \mathbb{R}$, $\lambda(h)$ denotes $\int_{\mathbb{R}^d} h(\theta) d\lambda(\theta)$ when it exists. Note that with such notations, for any $k \in \mathbb{N}^*$, probability measure λ on $\mathcal{B}(\mathbb{R}^d)$, measurable function $h : \mathbb{R}^d \to \mathbb{R}_+$, we have $\lambda(R_\gamma^k h) = (\lambda R_\gamma^k)(h)$. A probability measure π_γ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is said to be a invariant probability measure for R_γ, $\gamma > 0$, if $\pi_\gamma R_\gamma = R_\gamma$. A Markov chain $(\theta_k^{(\gamma)})_{k\in\mathbb{N}}$ satisfying the SGD recursion (1) for $\gamma > 0$ will be said at stationarity if it admits an invariant probability measure π_γ and $\theta_k^{(\gamma)}$ is distributed according to π_γ. Note that in this case for all $k \in \mathbb{N}$, the distribution of $\theta_k^{(\gamma)}$ is π_γ.

To show that $(\theta_k^{(\gamma)})_{k\geq 0}$ admits a unique stationary distribution π_γ and quantify the convergence of $(\nu_0 R_\gamma^k)_{k\geq 0}$ to π_γ, we use the Wasserstein distance, see [58]. A probability measure λ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is said to have a finite second moment if $\int_{\mathbb{R}^d} \|\vartheta\|^2 \lambda(\mathrm{d}\vartheta) < +\infty$. The set of probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ having a finite second moment is denoted by $\mathcal{P}_2(\mathbb{R}^d)$. For all probability measures ν and λ in $\mathcal{P}_2(\mathbb{R}^d)$, define the *Wasserstein distance* of order 2 between λ and ν by

$$W_2(\lambda, \nu) = \inf_{\xi \in \Pi(\lambda,\nu)} \Big(\int \|x-y\|^2 \xi(dx, dy) \Big)^{1/2} ,$$

where $\Pi(\mu, \nu)$ is the set of probability measure ξ on $\mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d)$ satisfying for all $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $\xi(\mathsf{A} \times \mathbb{R}^d) = \nu(\mathsf{A})$, $\xi(\mathbb{R}^d \times \mathsf{A}) = \lambda(\mathsf{A})$.

PROPOSITION 2. *Assume **A**1-**A**2-**A**3-**A**4(2). For any step-size $\gamma \in (0, 2/L)$, the Markov chain $(\theta_k^{(\gamma)})_{k\geq 0}$, defined by the recursion (1), admits a unique stationary distribution $\pi_\gamma \in \mathcal{P}_2(\mathbb{R}^d)$. In addition*

(a) for all $\theta \in \mathbb{R}^d$, $k \in \mathbb{N}^*$:

$$W_2^2(R_\gamma^k(\theta, \cdot), \pi_\gamma) \leq (1 - 2\mu\gamma(1 - \gamma L/2))^k \int_{\mathbb{R}^d} \|\theta - \vartheta\|^2 \,\mathrm{d}\pi_\gamma(\vartheta) ;$$

(b) for any Lipschitz function $\phi : \mathbb{R}^d \to \mathbb{R}$, with Lipschitz constant L_ϕ, for all $\theta \in \mathbb{R}^d$, $k \in \mathbb{N}^*$:

$$\left| R_\gamma^k \phi(\theta) - \pi_\gamma(\phi) \right| \leq L_\phi (1 - 2\mu\gamma(1 - \gamma L/2))^{k/2} \left(\int \|\theta - \vartheta\|^2 \mathrm{d}\pi_\gamma(\vartheta) \right)^{1/2} .$$

PROOF. Let $\gamma \in (0, 2/L)$ and $\lambda_1, \lambda_2 \in \mathcal{P}_2(\mathbb{R}^d)$. By [58, Theorem 4.1], there exists a couple of random variables $\theta_0^{(1)}, \theta_0^{(2)}$ such that $W_2^2(\lambda_1, \lambda_2) = \mathbb{E}[\|\theta_0^{(1)} - \theta_0^{(2)}\|^2]$ independent of $(\varepsilon_k)_{k\in\mathbb{N}^*}$. Let $(\theta_k^{(1)})_{k\geq 0}$,$(\theta_k^{(2)})_{k\geq 0}$ be the SGD iterates associated with the step-size γ, starting from $\theta_0^{(1)}$ and $\theta_0^{(2)}$ respectively and sharing the same noise, *i.e.* for all $k \geq 0$,

$$\begin{cases} \theta_{k+1}^{(1)} &= \theta_k^{(1)} - \gamma\big[f'(\theta_k^{(1)}) + \varepsilon_{k+1}(\theta_k^{(1)})\big] \\ \theta_{k+1}^{(2)} &= \theta_k^{(2)} - \gamma\big[f'(\theta_k^{(2)}) + \varepsilon_{k+1}(\theta_k^{(2)})\big] . \end{cases} \tag{11}$$

Note that using that $\theta_0^{(1)}, \theta_0^{(2)}$ are independent of ε_1, we have for $i, j \in \{1, 2\}$ using **A**3, that

$$\mathbb{E}[\langle \theta_0^{(i)}, \varepsilon(\theta_0^{(j)}) \rangle] = 0 . \tag{12}$$

Since for all $k \geq 0$, the distribution of $(\theta_k^{(1)}, \theta_k^{(2)})$ belongs to $\Pi(\lambda_1 R_\gamma^k, \lambda_2 R_\gamma^k)$, by definition of the Wasserstein distance we get

$$
\begin{aligned}
&W_2^2(\lambda_1 R_\gamma, \lambda_2 R_\gamma) \leq \mathbb{E}\left[\|\theta_1^{(1)} - \theta_1^{(2)}\|^2\right] \\
&= \mathbb{E}\left[\|\theta_0^{(1)} - \gamma f_1'(\theta_0^{(1)}) - (\theta_0^{(2)} - \gamma f_1'(\theta_0^{(2)})))\|^2\right] \\
&\overset{i)}{=} \mathbb{E}\left[\left\|\theta_0^{(1)} - \theta_0^{(2)}\right\|^2 - 2\gamma\left\langle f'(\theta_0^{(1)}) - f'(\theta_0^{(2)}), \theta_0^{(1)} - \theta_0^{(2)}\right\rangle\right] \\
&\qquad\qquad\qquad\qquad + \gamma^2 \mathbb{E}\left[\left\|f_1'(\theta_0^{(1)}) - f_1'(\theta_0^{(2)})\right\|^2\right] \\
&\overset{ii)}{\leq} \mathbb{E}\left[\left\|\theta_0^{(1)} - \theta_0^{(2)}\right\|^2 - 2\gamma(1-\gamma L/2)\left\langle f'(\theta_0^{(1)}) - f'(\theta_0^{(2)}), \theta_0^{(1)} - \theta_0^{(2)}\right\rangle\right] \\
&\overset{iii)}{\leq} (1 - 2\mu\gamma(1-\gamma L/2))\mathbb{E}\left[\left\|\theta_0^{(1)} - \theta_0^{(2)}\right\|^2\right] ,
\end{aligned}
$$

using (12) for $i)$, **A**4(2) for $ii)$, and finally **A**1 for $iii)$.

Thus by a straightforward induction, we get, setting $\rho = (1 - 2\mu\gamma(1-\gamma L/2))$

$$
\begin{aligned}
W_2^2(\lambda_1 R_\gamma^k, \lambda_2 R_\gamma^k) &\leq \mathbb{E}\left[\|\theta_k^{(1)} - \theta_k^{(2)}\|^2\right] \\
&\leq \rho \mathbb{E}\left[\|\theta_{k-1}^{(1)} - \theta_{k-1}^{(2)}\|^2\right] \leq \rho^k W_2^2(\lambda_1, \lambda_2) . \qquad (13)
\end{aligned}
$$

Since by **A**2-**A**3-**A**4(2), $\lambda_1 R_\gamma \in \mathcal{P}_2(\mathbb{R}^d)$, taking $\lambda_2 = \lambda_1 R_\gamma$ in (13), for any $N \in \mathbb{N}^*$, we have $\sum_{k=1}^N W_2^2(\lambda_1 R_\gamma^k, \lambda_2 R_\gamma^k) \leq \sum_{k=1}^N \rho^k W_2^2(\lambda_1, \lambda_1 R_\gamma)$. Therefore, we get $\sum_{k=1}^{+\infty} W_2^2(\lambda_1 R_\gamma^k, \lambda_1 R_\gamma^{k+1}) < +\infty$. By [58, Theorem 6.16], the space $\mathcal{P}_2(\mathbb{R}^d)$ endowed with W_2 is a Polish space. Then, $(\lambda_1 R_\gamma^k)_{k \geq 0}$ is a Cauchy sequence and converges to a limit $\pi_\gamma^{\lambda_1} \in \mathcal{P}_2(\mathbb{R}^d)$:

$$
\lim_{k \to +\infty} W_2(\lambda_1 R_\gamma^k, \pi_\gamma^{\lambda_1}) = 0 . \qquad (14)
$$

We show that the limit $\pi_\gamma^{\lambda_1}$ does not depend on λ_1. Assume that there exists $\pi_\gamma^{\lambda_2}$ such that $\lim_{k \to +\infty} W_2(\lambda_2 R_\gamma^k, \pi_\gamma^{\lambda_2}) = 0$. By the triangle inequality

$$
W_2(\pi_\gamma^{\lambda_1}, \pi_\gamma^{\lambda_2}) \leq W_2(\pi_\gamma^{\lambda_1}, \lambda_1 R_\gamma^k) + W_2(\lambda_1 R_\gamma^k, \lambda_2 R_\gamma^k) + W_2(\pi_\gamma^{\lambda_2}, \lambda_2 R_\gamma^k) .
$$

Thus by (13) and (14), taking the limits as $k \to +\infty$, we get $W_2(\pi_\gamma^{\lambda_1}, \pi_\gamma^{\lambda_2}) = 0$ and $\pi_\gamma^{\lambda_1} = \pi_\gamma^{\lambda_2}$. The limit is thus the same for all initial distributions and is denoted by π_γ.

Moreover, π_γ is invariant for R_γ. Indeed for all $k \in \mathbb{N}^*$,

$$
W_2(\pi_\gamma R_\gamma, \pi_\gamma) \leq W_2(\pi_\gamma R_\gamma, \pi_\gamma R_\gamma^k) + W_2(\pi_\gamma R_\gamma^k, \pi_\gamma) .
$$

Using (13) and (14), we get taking $k \to +\infty$, $W_2(\pi_\gamma R_\gamma, \pi_\gamma) = 0$ and $\pi_\gamma R_\gamma = \pi_\gamma$. The fact that π_γ is the unique stationary distribution is straightforward by contradiction and using (13).

Taking $\lambda_1 = \delta_\theta$, $\lambda_2 = \pi_\gamma$, using the invariance of π_γ and (13), we get (a). Finally, note that $\int_{\mathbb{R}^d} \|\theta - \vartheta\|^2 \mathrm{d}\pi_\gamma(\vartheta) < +\infty$ follows from the inequality for $a, b \in \mathbb{R}^d$, $\|a-b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$ and since we have established that $\pi_\gamma \in \mathcal{P}_2(\mathbb{R}^d)$.

Finally, if we take $\lambda_1 = \delta_\theta$ and $\lambda_2 = \pi_\gamma$, using $\pi_\gamma R_\gamma = \pi_\gamma$, (13), and the Cauchy-Schwarz inequality, we have for any $k \in \mathbb{N}^*$:

$$\begin{aligned}\left|R_\gamma^k \phi(\theta) - \pi_\gamma(\phi)\right| = \left|\mathbb{E}\left[\phi(\theta_{k,\gamma}^{(1)}) - \phi(\theta_{k,\gamma}^{(2)}))\right]\right| &\leq L_\phi \mathbb{E}^{1/2}[\left\|\theta_{k,\gamma}^{(1)} - \theta_{k,\gamma}^{(2)}\right\|^2] \\ &\leq L_\phi(1 - 2\mu\gamma(1-\gamma L/2))^{k/2}\left(\int \|\theta - \vartheta\|^2 \mathrm{d}\pi_\gamma(\vartheta)\right)^{1/2} ,\end{aligned}$$

which concludes the proof of (b).

□

A consequence of Proposition 2 is that the expectation of $\bar{\theta}_k^{(\gamma)}$ defined by (2) converges to $\int_{\mathbb{R}^d} \vartheta \mathrm{d}\pi_\gamma(\vartheta)$ as k goes to infinity at a rate of order $O(k^{-1})$, see Proposition 16 in Section 6.2.

3.1. *Expansion of moments of π_γ when γ is in a neighborhood of 0.* In this sub-section, we analyze the properties of the chain starting at θ_0 distributed according to π_γ. As a result, we prove that the mean of the stationary distribution $\bar{\theta}_\gamma = \int_{\mathbb{R}^d} \vartheta \pi_\gamma(\mathrm{d}\vartheta)$ is such that $\bar{\theta}_\gamma = \theta^* + \gamma\Delta + O(\gamma^2)$. Simple developments of Equation (1) at equilibrium result in expansions of the first two moments of the chain. It extends [45, 32] which showed that $(\gamma^{-1/2}(\pi_\gamma - \delta_{\theta^*}))_{\gamma>0}$ converges in distribution to a normal law as $\gamma \to 0$.

Quadratic case. When f is a quadratic function, *i.e.* f' is affine, we have the following result.

PROPOSITION 3. *Assume $f = f_\Sigma$, $f_\Sigma : \theta \mapsto \left\|\Sigma^{1/2}(\theta - \theta^*)\right\|^2/2$, where Σ is a positive definite matrix, and **A**2-**A**3-**A**4(4). Let $\gamma \in (0, 2/L)$. Then, it holds $\bar{\theta}_\gamma = \theta^*$, $\Sigma \otimes I + I \otimes \Sigma - \gamma \Sigma \otimes \Sigma$ is invertible, and*

$$\int_{\mathbb{R}^d} (\theta - \theta^*)^{\otimes 2} \pi_\gamma(\mathrm{d}\theta) = \gamma(\Sigma \otimes I + I \otimes \Sigma - \gamma\Sigma \otimes \Sigma)^{-1}\left[\int_{\mathbb{R}^d} \mathcal{C}(\theta)\pi_\gamma(\mathrm{d}\theta)\right] ,$$

where $\bar{\theta}_\gamma$ and $\mathcal{C}$ are given by (3) and (5) respectively, and π_γ is the invariant probability measure of R_γ given by Proposition 2.

The first part of the result, which highlights the crucial fact that for a quadratic function, the mean under the limit distribution is the optimal point, is easy to prove. Indeed, since π_γ is invariant for $(\theta_k^{(\gamma)})_{k\geq 0}$, if $\theta_0^{(\gamma)}$ is distributed according to π_γ, then $\theta_1^{(\gamma)}$ is distributed according to π_γ as well. Thus as $\theta_1^{(\gamma)} = \theta_0^{(\gamma)} - \gamma f'(\theta_0^{(\gamma)}) + \gamma\varepsilon_1(\theta_0^{(\gamma)})$ taking expectations on both sides, we get $\int_{\mathbb{R}^d} f'(\vartheta)\mathrm{d}\pi_\gamma(\vartheta) = 0$. For a quadratic function, whose gradient is affine: $\int_{\mathbb{R}^d} f'(\vartheta)\mathrm{d}\pi_\gamma(\vartheta) = f'(\bar\theta_\gamma) = 0$ and thus $\bar\theta_\gamma = \theta^*$. This implies that the averaged iterate converges to θ^*, see, *e.g.* [5]. The proof for the second expression is given in Section 6.3.

General case. While the quadratic case led to particularly simple expressions, in general, we can only get a first order development of these expectations as $\gamma \to 0$. Note that it improves on [45], which shows a similar expansion but with an error of order of $O(\gamma^{3/2})$.

THEOREM 4. *Assume* ***A****1*-***A****2*-***A****3*-***A****4*$(6 \vee [2(k_\varepsilon + 1)])$-***A****5* *and let* $\gamma \in (0, 2/L)$. *Then* $f''(\theta^*) \otimes I + I \otimes f''(\theta^*)$ *is invertible and*

$$\bar\theta_\gamma - \theta^* = \gamma f''(\theta^*)^{-1} f'''(\theta^*)\mathbf{A}\mathcal{C}(\theta^*) + O(\gamma^2) \tag{15}$$

$$\int_{\mathbb{R}^d} (\theta - \theta^*)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta) = \gamma \mathbf{A}\mathcal{C}(\theta^*) + O(\gamma^2)\ , \tag{16}$$

where

$$\mathbf{A} = \left(f''(\theta^*) \otimes I + I \otimes f''(\theta^*)\right)^{-1}\ , \tag{17}$$

$\bar\theta_\gamma$ *and* $\mathcal{C}$ *are given by* (3) *and* (5) *respectively, and* π_γ *is the invariant probability measure of* R_γ *given by Proposition 2.*

PROOF. The proof is postponed to Section 6.4. □

This shows that $\gamma \mapsto \bar\theta_\gamma$ is a differentiable function at $\gamma = 0$. The "drift" $\bar\theta_\gamma - \theta^*$ can be understood as an additional error occurring because the function is non quadratic ($f'''(\theta^*) \neq 0$) and the step-sizes are not decaying to zero. The mean under the limit distribution is at distance γ from θ^*. In comparison, the final iterate oscillates in a sphere of radius proportional to $\sqrt{\gamma}$.

3.2. *Expansion for a given $\gamma > 0$ when k tends to $+\infty$.* In this subsection, we analyze the convergence of $\bar\theta_k^{(\gamma)}$ to $\bar\theta_\gamma$, when $k \to \infty$, and the convergence of $\mathbb{E}[\|\bar\theta_k^{(\gamma)} - \bar\theta_\gamma\|^2]$ to 0. Under suitable conditions [24], $\bar\theta_k^{(\gamma)}$ satisfies a central limit theorem: $\{\sqrt{k}(\bar\theta_k^{(\gamma)} - \bar\theta_\gamma)\}_{k\in\mathbb{N}^*}$ converges in law to a

d-dimensional Gaussian distribution with zero-mean. However, this result is purely asymptotic and we propose a new tighter development that describes how the initial conditions are forgotten. We show that the convergence behaves similarly to the convergence in the quadratic case, where the expected squared distance decomposes as a sum of a bias term, that scales as k^{-2}, and a variance term, that scales as k^{-1}, plus linearly decaying residual terms. We also describe how the asymptotic bias and variance can be easily expressed as moments of solutions associated with several *Poisson equations*.

For any Lipschitz function $\varphi : \mathbb{R}^d \to \mathbb{R}^q$, by Lemma 8 in Section 6.2, the function $\psi_\gamma = \sum_{i=0}^{+\infty} \{R_\gamma^i \varphi - \pi_\gamma(\varphi)\}$ is well-defined, Lipschitz and satisfies $\pi_\gamma(\psi_\gamma) = 0$, $(\mathrm{Id} - R_\gamma)\psi_\gamma = \varphi$. The function ψ_γ will be referred to as the *Poisson solution* associated with φ. Consider the three following functions:

- ψ_γ the Poisson solution associated with $\varphi : \theta \mapsto \theta - \theta^*$,
- ϖ_γ the Poisson solution associated with $\theta \mapsto \psi_\gamma(\theta)$,
- χ_γ^1 the Poisson solution associated with $\theta \mapsto (\psi_\gamma(\theta))^{\otimes 2}$,
- χ_γ^2 the Poisson solution associated with $\theta \mapsto ((\psi_\gamma - \varphi)(\theta))^{\otimes 2}$.

THEOREM 5. *Assume* **A***1*-**A***2*-**A***3*-**A***4(4) and let $\gamma \in (0, 1/(2L))$. Then setting $\rho = (1 - \gamma\mu)^{1/2}$, for any starting point $\theta_0 \in \mathbb{R}^d$, $k \in \mathbb{N}^*$,*

$$
\begin{aligned}
&\mathbb{E}\left[\bar{\theta}_k^{(\gamma)} - \bar{\theta}_\gamma\right] = k^{-1}(\psi_\gamma(\theta_0) + O(\rho^k)) , \\
&\mathbb{E}\left[\left(\bar{\theta}_k^{(\gamma)} - \bar{\theta}_\gamma\right)^{\otimes 2}\right] = k^{-1}\, \pi_\gamma \left(\psi_\gamma^{\otimes 2} - (\psi_\gamma - \varphi)^{\otimes 2}\right) \\
&\qquad\qquad - k^{-2}\left[\pi_\gamma\left(\varpi_\gamma \varphi^\top + \varphi \varpi_\gamma^\top\right) + \chi_\gamma^2(\theta_0) - \chi_\gamma^1(\theta_0)\right] + O(k^{-3}) ,
\end{aligned}
$$

where $\bar{\theta}_k^{(\gamma)}$, $\bar{\theta}_\gamma$ are given by (2) and (3) respectively, and π_γ is the invariant probability measure of R_γ given by Proposition 2.

Equation (5) is a sum of three terms: (i) a variance term, that scales as $1/k$, and does not depend on the initial distribution (but only on the asymptotic distribution π_γ), and (ii) a bias term, which scales as $1/k^2$, and depends on the initial point $\theta_0 \in \mathbb{R}^d$, (iii) a non-positive residual term, which scales as $1/k^2$.

PROOF. In order to give the intuition of the proof and to underline how the associated Poisson solutions are introduced, we here sketch the proof of the first result. By definition of $\varphi : \theta \mapsto \theta - \theta^*$ and since ψ_γ satisfies $(\mathrm{Id} - R_\gamma)\psi_\gamma = \varphi$, we have

$$
\mathbb{E}\left[\bar{\theta}_{k+1}^{(\gamma)}\right] - \theta^* = (k+1)^{-1} \sum_{i=0}^{k} (R_\gamma^i \varphi)(\theta_0) = \pi_\gamma(\varphi) + (k+1)^{-1}\psi_\gamma(\theta_0) + R_\gamma^{k+1}\psi_\gamma(\theta_0),
$$

where we have used that

$$\sum_{i=0}^{\infty} R_\gamma^i(\varphi - \pi_\gamma(\varphi)) - R_\gamma^{k+1} \sum_{i=0}^{\infty} R_\gamma^i(\varphi - \pi_\gamma(\varphi)) = \psi_\gamma - R_\gamma^{k+1}\psi_\gamma \ .$$

Finally, we have that $R_\gamma^k \psi_\gamma(\theta_0)$ converges to 0 at linear speed, using Proposition 2 and $\pi_\gamma(\psi_\gamma) = 0$.

The formal and complete proof of this result is postponed to Section 6.5. □

This result gives an exact closed form for the asymptotic bias and variance, for a fixed γ, as $k \to \infty$. Unfortunately, in the general case, it is neither possible to compute the Poisson solutions exactly, nor is it possible to prove a first order development of the limits as $\gamma \to 0$.

When f_Σ is a quadratic function, it is possible, for any $\gamma > 0$, to compute ψ_γ and $\chi_\gamma^{1,2}$ explicitly; we get the following decomposition of the error, which exactly recovers the result of [15].

COROLLARY 6. *Assume that f is an objective function of a least-square regression problem, i.e. with the notations of Example 1, $f = f_\Sigma$, $\Sigma = \mathbb{E}[XX^\top]$, ε_k are defined by (6), and step-size $\gamma \leq 1/r^2$, with r defined by (8). Assume* ***A****1-****A****2-****A****3-****A****4(4). For any starting point $\theta_0 \in \mathbb{R}^d$:*

$$\begin{aligned}
\mathbb{E}\bar{\theta}_k^{(\gamma)} - \theta^* &= (1/(k\gamma))\Sigma^{-1}(\theta_0 - \theta^*) + O(\rho^k) \\
\mathbb{E}\left[\left(\bar{\theta}_k^{(\gamma)} - \theta^*\right)^{\otimes 2}\right] &= (1/k)\Sigma^{-1}\left\{\int_{\mathbb{R}^d} \mathcal{C}(\theta)\mathrm{d}\pi_\gamma(\theta)\right\}\Sigma^{-1} \\
&\quad + (1/(k^2\gamma^2))\Sigma^{-1}\Omega\left[\varphi(\theta_0)^{\otimes 2} - \pi_\gamma(\varphi^{\otimes 2})\right]\Sigma^{-1} \\
&\quad - (1/(k^2\gamma^2))(\Sigma^{-2}\otimes \mathrm{Id} + \mathrm{Id}\otimes\Sigma^{-2})\pi_\gamma(\varphi^{\otimes 2}) + O(k^{-3}) \ .
\end{aligned}$$

With $\Omega = (\Sigma\otimes I + I\otimes\Sigma - \gamma\Sigma\otimes\Sigma)(\Sigma\otimes I + I\otimes\Sigma - \gamma\mathbf{T})^{-1}$, and

$$\mathbf{T} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d} \ , \ A \mapsto \mathbb{E}\left[(X^\top A X)XX^\top\right] \ . \tag{18}$$

PROOF. The proof is postponed to the supplementary paper [17, Section S5]. □

The bound on the second order moment is composed of a variance term $k^{-1}\Sigma^{-1}\pi_\gamma(\mathcal{C})\Sigma^{-1}$, a bias term which decays as k^{-2}, and a non-positive residual term. Note that the bias is 0 if we start under the limit distribution.

3.3. *Continuous interpretation of SGD and weak error expansion.* Under the stated assumptions on f and $(\varepsilon_k)_{k\in\mathbb{N}^*}$, we have analyzed the convergence of the stochastic gradient recursion (1). We here describe how this recursion can be seen as a noisy discretization of the following *gradient flow* equation, for $t \in \mathbb{R}_+$:

$$\dot{\theta}_t = -f'(\theta_t) \; . \tag{19}$$

Note that since $f'(\theta^*) = 0$ by definition of θ^* and **A**1, then θ^* is an equilibrium point of (19), *i.e.* $\theta_t = \theta^*$ for all $t \geq 0$ if $\theta_0 = \theta^*$. Under **A**2, (19) admits a unique solution on $\mathbb{R}_+$ for any starting point $\theta \in \mathbb{R}^d$. Denote by $(\varphi_t)_{t\geq 0}$ the flow of (19), defined for all $\theta \in \mathbb{R}^d$ by $(\varphi_t(\theta))_{t\geq 0}$ as the solution of (19) starting at θ.

Denote by $(\mathcal{A}, D(\mathcal{A}))$, the *infinitesimal generator* associated with the flow $(\varphi_t)_{t\geq 0}$ defined by

$$\mathrm{D}(\mathcal{A}) = \left\{ h : \mathbb{R}^d \to \mathbb{R} \; : \; \text{ for all } \theta \in \mathbb{R}^d, \; \lim_{t\to 0} \frac{h(\varphi_t(\theta)) - h(\theta)}{t} \text{ exists } \right\}$$

$$\mathcal{A}h(\theta) = \lim_{t\to 0} \frac{\{h(\varphi_t(\theta)) - h(\theta)\}}{t} \text{ for all } h \in \mathrm{D}(\mathcal{A}) \; , \quad \theta \in \mathbb{R}^d \; . \tag{20}$$

Note that for any $h \in \mathrm{C}^1(\mathbb{R}^d)$, $h \in D(\mathcal{A})$, $\mathcal{A}h = -\langle f', h'\rangle$.

Under **A**1 and **A**2, for any locally Lipschitz function $g : \mathbb{R}^d \to \mathbb{R}$ (extension to a function $g : \mathbb{R}^d \to \mathbb{R}^q$ can easily be done considering all assumptions and results coordinatewise), denote by h_g the solution of the continuous Poisson equation defined for all $\theta \in \mathbb{R}^d$ by $h_g(\theta) = \int_0^\infty (g(\varphi_s(\theta)) - g(\theta^*))\mathrm{d}s$. Note that h_g is well-defined by Lemma 21-b) in Section 6.6.1, since g is assumed to be locally Lipschitz. Roughly, Lemma 21-b) implies that for any $\theta \in \mathbb{R}^d$, there exists $C(\theta) \geq 0$ such that for any $s \in \mathbb{R}_+$, $|g(\varphi_s(\theta)) - g(\theta^*)| \leq C(\theta)\mathrm{e}^{-s}$, and therefore $s \mapsto g(\varphi_s(\theta)) - g(\theta^*)$ is integrable on $\mathbb{R}_+$ for any $\theta \in \mathbb{R}^d$. By (20), we have for all $g : \mathbb{R}^d \to \mathbb{R}$, locally Lipschitz,

$$\mathcal{A}h_g(\theta) = g(\theta^*) - g(\theta) \; . \tag{21}$$

Under regularity assumptions on g (see Theorem 23), h_g is continuously differentiable and therefore satisfies $\langle f', h_g'\rangle = g - g(\theta^*)$. The idea is then to make a Taylor expansion of $h_g(\theta_{k+1}^{(\gamma)})$ around $\theta_k^{(\gamma)}$ to express $k^{-1}\sum_{i=1}^k g(\theta_i^{(\gamma)}) - g(\theta^*)$ as convergent terms involving the derivatives of h_g. For $g : \mathbb{R}^d \to \mathbb{R}$ and $\ell, p \in \mathbb{N}$, $\ell \geq 1$ consider the following assumptions.

A6 (ℓ, p). *There exist $a_g, b_g \in \mathbb{R}_+$ such that $g \in \mathrm{C}^\ell(\mathbb{R}^d)$ and for all $\theta \in \mathbb{R}^d$ and $i \in \{1, \cdots, \ell\}$, $\left\|g^{(i)}(\theta)\right\| \leq a_g \{\|\theta - \theta^*\|^p + b_g\}$.*

THEOREM 7. *Let $g : \mathbb{R}^d \to \mathbb{R}$ satisfying* **A**6$(5, p)$ *for $p \in \mathbb{N}$. Assume* **A**1-**A**2-**A**3-**A**5*. Furthermore, suppose that there exists $q \in \mathbb{N}$ and $C \geq 0$ such that for all $\theta \in \mathbb{R}^d$,*

$$\mathbb{E}\left[\|\varepsilon_1(\theta)\|^{p+k_\varepsilon+3}\right] \leq C(1 + \|\theta - \theta^*\|^q) ,$$

and **A**4$(2\tilde{p})$ *holds for $\tilde{p} = p + 3 + q \vee k_\varepsilon$. Then there exists a constant $\varsigma > 0$ only depending on $\tilde{p}$ such that for all $\gamma \in (0, 1/(\varsigma L))$, $k \in \mathbb{N}^*$ and any starting point $\theta_0 \in \mathbb{R}^d$ it holds that:*

$$\mathbb{E}\left[k^{-1}\sum_{i=1}^{k}\left\{g(\theta_i^{(\gamma)}) - g(\theta^*)\right\}\right] = (1/(k\gamma))\left\{h_g(\theta_0) - \mathbb{E}\left[h_g(\theta_{k+1}^{(\gamma)})\right]\right\} + (\gamma/2)\operatorname{tr}\left(h_g''(\theta^*)\,\mathcal{C}(\theta^*)\right) - (\gamma/k)A_1(\theta_0) - \gamma^2 A_2(\theta_0, k) , \tag{22}$$

where $\theta_k^{(\gamma)}$ is the Markov chain starting from θ_0 and defined by the recursion (1) and $\mathcal{C}$ is given by (5). In addition for some constant $C \geq 0$ independent of γ and k, we have

$$A_1(\theta_0) \leq C\left\{1 + \|\theta_0 - \theta^*\|^{\tilde{p}}\right\} , \; A_2(\theta_0, k) \leq C\left\{1 + \|\theta_0 - \theta^*\|^{\tilde{p}}/k\right\} .$$

PROOF. The proof is postponed to Section 6.6. □

First in the case where f' is affine, choosing for g the identity function, then $h_{\mathrm{Id}} = \int_0^{+\infty}\{\varphi_s - \theta^*\}\mathrm{d}s = \Sigma^{-1}$, and we get that the first term in (22) vanishes which is expected since in that case $\bar{\theta}_\gamma = \theta^*$. Second by Lemma 22-b), we recover the first expansion of Theorem 4 for arbitrary objective functions f. Finally note that for all $q \in \mathbb{N}$, under appropriate conditions, Theorem 7 implies that there exist constants $C_1, C_2(\theta_0) \geq 0$ such that $\mathbb{E}\left[k^{-1}\sum_{i=1}^{k}\|\theta_i^{(\gamma)} - \theta^*\|^{2q}\right] = C_1\gamma + C_2(\theta_0)/k + O(\gamma^2)$.

3.4. *Discussion.* Classical proofs of convergence rely on another decomposition, originally proposed by [43] and used in recent papers analyzing the averaged iterate [4] . We here sketch the arguments of these decompositions, in order to highlight the main difference, namely the fact that the residual term is not well controlled when γ goes to zero in the classical proof.

Classical decomposition. The starting point of this decomposition is to consider a Taylor expansion of $f'(\theta_{k+1}^{(\gamma)})$ around θ^*. For any $k \in \mathbb{N}$,

$$f'(\theta_k^{(\gamma)}) = f''(\theta^*)(\theta_k^{(\gamma)} - \theta^*) + O\left(\left\|\theta_k^{(\gamma)} - \theta^*\right\|^2\right) .$$

As a consequence, using the definition of the SGD recursion (1),

$$\begin{aligned}\theta_{k+1}^{(\gamma)} - \theta_k^{(\gamma)} &= -\gamma f'(\theta_k^{(\gamma)}) - \gamma\varepsilon_{k+1}(\theta_k^{(\gamma)}) \\ &= -\gamma f''(\theta^*)(\theta_k^{(\gamma)} - \theta^*) - \gamma\varepsilon_{k+1}(\theta_k^{(\gamma)}) + \gamma O\left(\left\|\theta_k^{(\gamma)} - \theta^*\right\|^2\right) .\end{aligned}$$

Thus

$$f''(\theta^*)(\theta_k^{(\gamma)} - \theta^*) = \gamma^{-1}(-\theta_{k+1}^{(\gamma)} + \theta_k^{(\gamma)}) - \varepsilon_{k+1}(\theta_k^{(\gamma)}) + O\left(\left\|\theta_k^{(\gamma)} - \theta^*\right\|^2\right) .$$

Averaging over the first k iterates yields:

$$\begin{aligned}(k+1)\left(\bar{\theta}_k^{(\gamma)} - \theta^*\right) &= \gamma^{-1} f''(\theta^*)^{-1}\left(\theta_0^{(\gamma)} - \theta_{k+1}^{(\gamma)}\right) - \sum_{i=0}^{k} f''(\theta^*)^{-1}\varepsilon_{i+1}\left(\theta_i^{(\gamma)}\right) \\ &\quad + \sum_{i=0}^{k} O\left(\left\|\theta_i^{(\gamma)} - \theta^*\right\|^2\right). \end{aligned} \tag{23}$$

The term on the right-hand part of Equation (23) is composed of a bias term (depending on the initial condition), a variance term, and a residual term. This residual term differentiates the general setting from the quadratic one (in which it does not appear, as the first-order Taylor expansion of f' is exact). This decomposition has been used in [4] to prove upper bounds on the error, but does not allow for a tight decomposition in powers of γ when $\gamma \to 0$. Indeed, the residual $\theta_i^{(\gamma)} - \theta^*$ simply does not go to 0 when $\gamma \to 0$: on the contrary, the chain becomes ill-conditioned when $\gamma = 0$.

New decomposition. Here, we use the fact that for a function $g : \mathbb{R}^d \to \mathbb{R}^q$ regular enough, there exists $h_g : \mathbb{R}^d \to \mathbb{R}^q$ satisfying, for any $\theta \in \mathbb{R}^d$:

$$h_g'(\theta) f'(\theta) = g(\theta) - g(\theta^*),$$

where $h_g'(\theta) \in \mathbb{R}^{q\times d}$, and $f'(\theta) \in \mathbb{R}^d$. The starting point is then a first-order Taylor development of $h_g(\theta_{k+1}^{(\gamma)})$ around $\theta_k^{(\gamma)}$. For any $k \in \mathbb{N}^*$, we have

$$\begin{aligned}h_g(\theta_{k+1}^{(\gamma)}) &= h_g(\theta_k^{(\gamma)}) + h_g'(\theta_k^{(\gamma)})(\theta_{k+1}^{(\gamma)} - \theta_k^{(\gamma)}) + O\left(\left\|\theta_{k+1}^{(\gamma)} - \theta_k^{(\gamma)}\right\|^2\right) \\ &= h_g(\theta_k^{(\gamma)}) - \gamma h_g'(\theta_k^{(\gamma)}) f'(\theta_k^{(\gamma)}) - \gamma h_g'(\theta_k^{(\gamma)})\varepsilon_{k+1}(\theta_k^{(\gamma)}) + O\left(\left\|\theta_{k+1}^{(\gamma)} - \theta_k^{(\gamma)}\right\|^2\right) \\ &= h_g(\theta_k^{(\gamma)}) - \gamma(g(\theta_k^{(\gamma)}) - g(\theta^*)) - \gamma h_g'(\theta_k^{(\gamma)})\varepsilon_{k+1}(\theta_k^{(\gamma)}) + O\left(\left\|\theta_{k+1}^{(\gamma)} - \theta_k^{(\gamma)}\right\|^2\right).\end{aligned}$$

Thus reorganizing terms,

$$g(\theta_k^{(\gamma)}) - g(\theta^*) = \gamma^{-1}\left\{h_g(\theta_k^{(\gamma)}) - h_g(\theta_{k+1}^{(\gamma)})\right\} + h_g'(\theta_k^{(\gamma)})\varepsilon_{k+1}(\theta_k^{(\gamma)}) + \gamma^{-1}O\left(\left\|\theta_{k+1}^{(\gamma)} - \theta_k^{(\gamma)}\right\|^2\right).$$

Finally, averaging over the first k iterations and taking $g = \mathrm{Id}$ give

$$(k+1)\left(\bar{\theta}_k^{(\gamma)} - \theta^*\right) = \gamma^{-1}\left(h_{\mathrm{Id}}(\theta_0^{(\gamma)}) - h_{\mathrm{Id}}(\theta_{k+1}^{(\gamma)})\right) + \sum_{i=0}^{k} h_{\mathrm{Id}}'(\theta_i^{(\gamma)})\varepsilon_{i+1}\left(\theta_i^{(\gamma)}\right) + \gamma^{-1}\sum_{i=0}^{k} O\left(\left\|\theta_{i+1}^{(\gamma)} - \theta_i^{(\gamma)}\right\|^2\right). \tag{24}$$

This expansion is the root of the proof of Theorem 7, which formalizes the expansion as powers of γ. The key difference between decompositions (23) and (24) is that in the latter, when $\gamma \to 0$, the expectation of the residual term tends to 0 and can naturally be controlled.

4. Experiments. We performed experiments on simulated data, for logistic regression, with $n = 10^7$ observations, for $d = 12$ and 4. Results are presented in Figure 2. The data are a.s. bounded by $R \geq 0$, therefore $R^2 = L$. We consider SGD with constant step-sizes $1/R^2$, $1/2R^2$ (and $1/4R^2$) with or without averaging, with $R^2 = L$. Without averaging, the chain saturates with an error proportional to γ (since $\|\theta_k^{(\gamma)} - \theta^*\| = O(\sqrt{\gamma})$ as $k \to +\infty$). Note that the ratio between the convergence limits of the two sequences is roughly 2 in the un-averaged case, and 4 in the averaged case, which confirms the predicted limits. We consider Richardson Romberg iterates, which saturate at a much lower level, and performs much better than decaying step-sizes (as $1/\sqrt{k}$) on the first iterations, as it forgets the initial conditions faster. Finally, we run the online-Newton algorithm [5], which performs very well but has no convergence guarantee. On the right plot, we also propose an estimator that uses 3 different step-sizes to perform a higher order interpolation. More precisely, for all $k \in \mathbb{N}^*$, we compute $\tilde{\theta}_k^3 = \frac{8}{3}\bar{\theta}_k^{(\gamma)} - 2\bar{\theta}_k^{(2\gamma)} + \frac{1}{3}\bar{\theta}_k^{(4\gamma)}$. With such an estimator, the *first 2* terms in the expansion, scaling as γ and γ^2, should vanish, which explains why it does not saturate.

We also performed experiments on the covertype dataset (581012 observations, $d = 54$), obtained from the LIBSVM data website[3]. Similarly, Richardson Romberg iterates outperform constant step-size, while decaying step sizes are particularly slow. Convergence results are given in Figure 3.

[3]http://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets

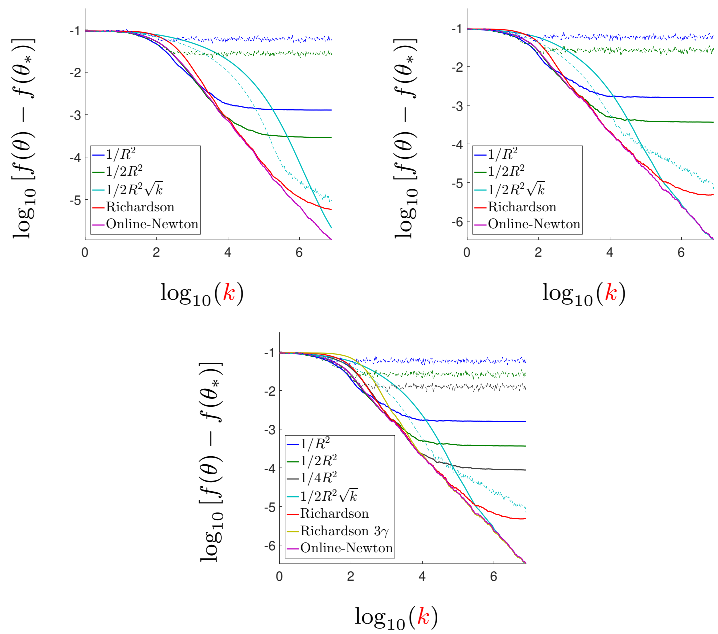

Figure 2: Synthetic data, logarithmic scales. Upper-left: logistic regression, $d = 12$, with averaged SGD with step-size $1/R^2$, $1/2R^2$, decaying step-sizes ($\gamma_k = 1/(2R^2\sqrt{k})$) (averaged (plain) and non-averaged (dashed)), Richardson-Romberg extrapolated iterates, and online Newton iterates. Upper-right: same in lower dimension ($d = 4$). Bottom: same but with three different step-sizes and an estimator built using the Richardson estimator $\tilde{\theta}_k^3 = \frac{8}{3}\bar{\theta}_k^{(\gamma)} - 2\bar{\theta}_k^{(2\gamma)} + \frac{1}{3}\bar{\theta}_k^{(4\gamma)}$, with 3 different step-sizes 3γ, 2γ and $\gamma = 1/4R^2$.

5. Conclusion. In this paper, we have used and developed Markov chain tools to analyze the behavior of constant step-size SGD, with a complete analysis of its convergence, outlining the effect of initial conditions, noise and step-sizes. For machine learning problems, this allows us to extend known results from least-squares to all loss functions. This analysis leads naturally to using Romberg-Richardson extrapolation, that provably improves the convergence behavior of the averaged SGD iterates. Our work opens up several avenues for future work: (a) show that Richardson-Romberg trick can be applied to the decreasing step-sizes setting, (b) study the extension of our results under self-concordance condition [3].

6. Postponed proofs.

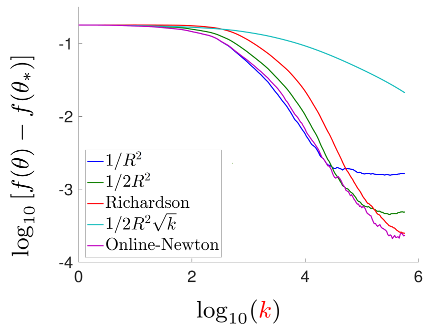

Figure 3: Covertype dataset.

6.1. *Discussion on assumptions on the noise.* Assumption **A**4, made in the text, can be weakened in order to apply to settings where input observations are un-bounded. Typically, Gaussian inputs would not satisfy Assumption **A**4: there exists no L such that almost surely f_k' is L-Lipschitz continuous and therefore Assumption **A**4 ($p = 2$) does not hold. Indeed, for least squares regression, as described in Example 1, we have $f_k'(\theta) - f_k'(\eta) = X_k X_k^\top(\theta - \eta)$, which is $||X_k||^2$-Lipshitz. If $||X_k||^2$ is not a.s. bounded, then there exists no L such that almost surely f_k' is L-Lipschitz.

However, in many cases, we only need Assumption **A**7 below. Let $p \geq 2$.

A7 (p). *(i) There exists $\tilde{\tau}_p \geq 0$ such that $\{\mathbb{E}^{1/p}[\|\varepsilon_1(\theta^*)\|^p]\} \leq \tilde{\tau}_p$.*
(ii) For all $x, y \in \mathbb{R}^d$, there exists $L \geq 0$ such that, for $q = 2, \dots, p$,

$$\mathbb{E}\left[\left\|f_1'(x) - f_1'(y)\right\|^q\right] \leq L^{q-1}\left\|x - y\right\|^{q-2}\left\langle x - y, f'(x) - f'(y)\right\rangle \,, \tag{25}$$

*where L is the same constant appearing in **A**2 and f_1' is defined by (4).*

For Gaussian inputs, assumption Assumption **A7** is satisfied, for example for A7($p = 2$): $\mathbb{E}||f_k'(\theta) - f_k'(\eta)||^2 = (\theta - \eta)^\top \mathbb{E}[||X_k||^2 X_k X_k^\top](\theta - \eta) \leq R^2(\theta - \eta)^\top \mathbb{E}[X_k X_k^\top](\theta - \eta)$.

On the other hand, we consider also the stronger assumption that the noise is independent of θ (referred to as the "semi-stochastic" setting, see [18]), or more generally that the noise has a uniformly bounded fourth order moment.

A8. *There exists $\tau \geq 0$ such that $\sup_{\theta \in \mathbb{R}^d}\{\mathbb{E}^{1/4}[\|\varepsilon_1(\theta)\|^4]\} \leq \tau$.*

Assumption **A**7(p), $p \geq 2$, is the weakest, as it is satisfied for random design least mean squares and logistic regression with bounded fourth moment of the inputs. Note that we do not assume that gradient or gradient estimates are a.s. bounded, to avoid the need for a constraint on the space where iterates live. It is straightforward to see that **A**7(p), $p \geq 2$, implies **A**4(p) with $\tau_p = \tilde{\tau}_p$, and **A**8-**A**2 implies **A**4(4).

It is important to note that assuming **A**3 (especially that $(\varepsilon_k)_{k\in\mathbb{N}^\star}$ are i.i.d. random fields) *does not* imply **A**8. On the contrary, making *the semi stochastic assumption*, *i.e.* that the noise functions $(\varepsilon_k(\theta_{k-1}))_{k\in\mathbb{N}^\star}$ are i.i.d. vectors (*e.g.* satisfied if ε_k is constant as a function of θ), is a very strong assumption, and implies **A**8.

Validity of the results under ***A**7(p).* Most of the results given in the main text would hold under **A**7(p), for p large enough. In the following proofs, we use **A**7 when possible. It is easy to see that, under, say **A**7($p = 10$), Propositions 2 and 3, Theorems 4 and 5 hold.

6.2. *Preliminary results.* We preface the proofs of the main results by some technical lemmas.

LEMMA 8. *Assume* ***A**1-**A**2-**A**3-**A**4(2). Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a L_ϕ-Lipschitz continuous function. For any step-size $\gamma \in (0, 2/L)$, the function $\psi_\gamma : \mathbb{R}^d \to \mathbb{R}$ defined for all $\theta \in \mathbb{R}^d$ by*

$$\psi_\gamma(\theta) = \sum_{i=0}^{+\infty} R_\gamma^i \phi(\theta) \ , \tag{26}$$

is well-defined, Lipschitz continuous and satisfies $(\mathrm{Id} - R_\gamma)\psi_\gamma = \phi$, $\pi_\gamma(\psi_\gamma) = 0$*. In addition, if* $\tilde{\psi}_\gamma : \mathbb{R}^d \to \mathbb{R}$ *is an other Lipchitz function satisfying* $(\mathrm{Id} - R_\gamma)\tilde{\psi}_\gamma = \phi$, $\pi_\gamma(\tilde{\psi}_\gamma) = 0$*, then* $\psi_\gamma = \tilde{\psi}_\gamma$.

PROOF. Let $\gamma \in (0, 2/L)$. By Proposition 2-(b), for any Lipschitz continuous function ϕ, $\{\theta \mapsto \sum_{i=1}^k (R_\gamma^i \phi(\theta) - \pi_\gamma(\phi))\}_{k\geq 0}$ converges absolutely on all compact sets of $\mathbb{R}^d$. Therefore ψ_γ given by (26) is well-defined. Let $(\theta, \vartheta) \in \mathbb{R}^d \times \mathbb{R}^d$. Consider now the two processes $(\theta_k^{(1)})_{\geq 0}$,$(\theta_k^{(2)})_{k\geq 0}$ defined by (11) with $\lambda_1 = \delta_\theta$ and $\lambda_2 = \delta_\vartheta$. Then, for any $k \in \mathbb{N}^*$, using (13):

$$\begin{aligned} \left| R_\gamma^k \phi(\theta) - R_\gamma^k \phi(\vartheta) \right| &\leq L_\phi \mathbb{E}^{1/2} \left[\left\| \theta_{k,\gamma}^{(1)} - \theta_{k,\gamma}^{(2)} \right\|^2 \right] \\ &\leq L_\phi (1 - 2\mu\gamma(1 - \gamma L/2))^{k/2} \|\theta - \vartheta\| \ . \end{aligned} \tag{27}$$

Therefore by definition (26), ψ_γ is Lipschitz-continuous. Finally, it is straightforward to verify that ψ_γ satisfies the stated properties.

If $\tilde{\psi}_\gamma : \mathbb{R}^d \to \mathbb{R}$ is an other Lipchitz function satisfying these properties, we have for all $\theta \in \mathbb{R}^d$, $(\psi_\gamma - \tilde{\psi}_\gamma)(\theta) = R_\gamma(\psi_\gamma - \tilde{\psi}_\gamma)(\theta)$. Therefore for all $k \in \mathbb{N}^*$, $\theta \in \mathbb{R}^d$, $(\psi_\gamma - \tilde{\psi}_\gamma)(\theta) = R_\gamma^k(\psi_\gamma - \tilde{\psi}_\gamma)(\theta)$. But by Proposition 2-(b), $\lim_{k\to+\infty} R_\gamma^k(\psi_\gamma - \tilde{\psi}_\gamma)(\theta) = \pi_\gamma(\psi_\gamma - \tilde{\psi}_\gamma) = 0$, which concludes the proof. $\square$

LEMMA 9. *Assume* ***A****1*-***A****2*-***A****3*-***A****4(2). Then we have for any $\gamma \in (0, 2/L)$.*

$$\int_{\mathbb{R}^d} f'(\theta)\pi_\gamma(\mathrm{d}\theta) = 0 \ .$$

PROOF. Let $(\theta_k^{(\gamma)})_{k\in\mathbb{N}}$ be a Markov chain satisfying (1), with $\theta_0^{(\gamma)}$ distributed according to π_γ. Then the proof follows from taking the expectation in (1) for $k = 0$, using that the distribution of $\theta_1^{(\gamma)}$ is π_γ, $\mathbb{E}[\varepsilon_1(\theta)] = 0$ for all $\theta \in \mathbb{R}^d$ and ε_1 is independent of $\theta_0^{(\gamma)}$. $\square$

LEMMA 10. *Assume* ***A****1*-***A****2*-***A****3*-***A****7(2). Then for any initial condition $\theta_0^{(\gamma)} \in \mathbb{R}^d$, we have for any $\gamma > 0$,*

$$\mathbb{E}\left[\left\|\theta_{k+1}^{(\gamma)} - \theta^*\right\|^2 \middle| \mathcal{F}_k\right] \le (1 - 2\gamma\mu(1-\gamma L))\left\|\theta_k^{(\gamma)} - \theta^*\right\|^2 + 2\gamma^2\tilde{\tau}_2^2 \ ,$$

where $(\theta_k^{(\gamma)})_{k\ge0}$ is given by (1). Moreover, if $\gamma \in (0, 1/L)$, we have

$$\int_{\mathbb{R}^d} \|\theta - \theta^*\|^2 \pi_\gamma(\mathrm{d}\theta) \le \gamma\tilde{\tau}_2^2/(\mu(1-\gamma L)) \ . \tag{28}$$

PROOF. The proof and result is very close to the ones from [41] but we extend it without a.s. Lipschitzness (**A**4) but with **A**7. Using **A**3-**A**1 and $f'(\theta^*) = 0$, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\theta_{k+1}^{(\gamma)} - \theta^*\right\|^2 \middle| \mathcal{F}_k\right] &\le \left\|\theta_k^{(\gamma)} - \theta^*\right\|^2 + \gamma^2\mathbb{E}\left[\left\|f'_{k+1}(\theta_k^{(\gamma)})\right\|^2 \middle| \mathcal{F}_k\right] \\
&\quad - 2\gamma\mathbb{E}\left[\left\langle f'_{k+1}(\theta_k^{(\gamma)}) - f'_{k+1}(\theta^*), \theta_k^{(\gamma)} - \theta^*\right\rangle \middle| \mathcal{F}_k\right] \quad (29) \\
&\le (1 - 2\mu\gamma)\left\|\theta_k^{(\gamma)} - \theta^*\right\|^2 + \gamma^2\mathbb{E}\left[\left\|f'_{k+1}(\theta_k^{(\gamma)})\right\|^2 \middle| \mathcal{F}_k\right] \ . \quad (30)
\end{aligned}$$

In addition, under **A**3-**A**7(2) and using (4), we have:

$$
\begin{aligned}
\mathbb{E}&\left[\left\|f'_{k+1}(\theta_k^{(\gamma)})\right\|^2\middle|\mathcal{F}_k\right] \\
&\leq 2\left(\mathbb{E}\left[\left\|f'_{k+1}(\theta_k^{(\gamma)})-f'_{k+1}(\theta^*)\right\|^2\middle|\mathcal{F}_k\right]+\mathbb{E}\left[\left\|f'_{k+1}(\theta^*)\right\|^2\middle|\mathcal{F}_k\right]\right) \\
&\leq 2\left(\mathbb{E}\left[\left\|f'_{k+1}(\theta_k^{(\gamma)})-f'_{k+1}(\theta^*)\right\|^2\middle|\mathcal{F}_k\right]+\tau^2\right) \\
&\leq 2\left(L\mathbb{E}\left[\left\langle f'_{k+1}(\theta_k^{(\gamma)})-f'_{k+1}(\theta^*),\theta_k^{(\gamma)}-\theta^*\right\rangle\middle|\mathcal{F}_k\right]+\tau^2\right) \\
&\leq 2\left(L\left\langle f'(\theta_k^{(\gamma)})-f'(\theta^*),\theta_k^{(\gamma)}-\theta^*\right\rangle+\tau^2\right) .
\end{aligned}
$$

Combining this result and (30) concludes the proof of the first inequality.

Regarding the second bound, let a fixed initial point $\theta_0^{(\gamma)} \in \mathbb{R}^d$. By Jensen inequality and the first result we get for any $k \in \mathbb{N}$ and $M \geq 0$,

$$
\begin{aligned}
\mathbb{E}\left[\left\|\theta_{k+1}^{(\gamma)}-\theta^*\right\|^2\wedge M\right] \leq (1-2\gamma\mu(1-\gamma L))^{k+1}\left\|\theta_0^{(\gamma)}-\theta^*\right\|^2 \\
+ 2\gamma^2\tilde{\tau}_2^2\sum_{i=0}^{k}(1-2\gamma\mu(1-\gamma L))^i .
\end{aligned}
$$

Since by Proposition 2-(b), $\lim_{k\to+\infty}\mathbb{E}[\|\theta_{k+1}^{(\gamma)}-\theta^*\|^2\wedge M] = \int_{\mathbb{R}^d}\{\|\theta-\theta^*\|^2 \wedge M\}\pi_\gamma(\mathrm{d}\theta)$, we get for any $M \geq 0$,

$$
\int_{\mathbb{R}^d}\{\|\theta-\theta^*\|^2\wedge M\}\pi_\gamma(\mathrm{d}\theta) \leq \gamma\tilde{\tau}_2^2/(\mu(1-\gamma L)) .
$$

Taking $M \to +\infty$ and applying the monotone convergence theorem concludes the proof. □

Using Lemma 10, we can extend Lemma 8 to functions ϕ which are locally Lipschitz.

LEMMA 11. *Assume* ***A****1*-***A****2*-***A****3*-***A****4(4). Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a function such that there exists $L_\phi \geq 0$ such that for any $x, y \in \mathbb{R}^d$,*

$$
|\phi(x)-\phi(y)| \leq L_\phi\|x-y\|\{1+\|x\|+\|y\|\} . \tag{31}
$$

For any step-size $\gamma \in (0, 1/L)$, it holds:

(a) there exists $C \geq 0$ such that for all $\theta \in \mathbb{R}^d$, $k \in \mathbb{N}^*$:

$$
\left|R_\gamma^k\phi(\theta)-\pi_\gamma(\phi)\right| \leq CL_\phi(1-2\mu\gamma(1-\gamma L))^{k/2}\left\{1+\|\theta-\theta^*\|^2\right\} ;
$$

(b) the function $\psi_\gamma : \mathbb{R}^d \to \mathbb{R}$ defined for all $\theta \in \mathbb{R}^d$ by (26) is well-defined satisfies $(\mathrm{Id} - R_\gamma)\psi_\gamma = \phi$, $\pi_\gamma(\psi_\gamma) = 0$ and there exists $L_\psi \geq 0$ such that for any $x, y \in \mathbb{R}^d$,

$$|\psi(x) - \psi(y)| \leq L_\psi \|x - y\| \{1 + \|x\| + \|y\|\} \; . \tag{32}$$

PROOF. The proof is similar to the proof of Proposition 2 (b) and Lemma 8. It is given in the supplementary paper [17, Section S1]. □

It is worth pointing out that under Assumption **A**8 (the "semi-stochastic" assumption), a slightly different result holds. The following result underlines the difference between a stochastic noise and a semi-stochastic noise, especially the fact that the maximal step-size differs depending on this assumption being made.

LEMMA 12. *Assume **A**1-**A**2-**A**3-**A**8. Then for any initial condition $\theta_0^{(\gamma)} \in \mathbb{R}^d$, we have for any $\gamma \in (0, 2/(m+L)]$,*

$$\mathbb{E}\left[\left\|\theta_{k+1}^{(\gamma)} - \theta^*\right\|^2 \middle| \mathcal{F}_k\right] \leq (1 - 2\gamma\mu L/(\mu + L)) \left\|\theta_k^{(\gamma)} - \theta^*\right\|^2 + \gamma^2\tau^2 \; ,$$

where $(\theta_k^{(\gamma)})_{k\geq 0}$ is given by (1).

PROOF. The proof is postponed to [17, Section S2]. □

We give uniform bounds on the moments of the chain $(\theta_k^{(\gamma)})_{k\geq 0}$ for $\gamma > 0$. For $p \geq 1$, recall that under **A**4(2p), the noise at optimal point has a moment of order $2p$ and we denote

$$\tau_{2p} = \mathbb{E}^{1/2p}\left[\|\varepsilon_1(\theta^*)\|^{2p}\right] \; . \tag{33}$$

We give a bound on the p-th order moment of the chain, under the assumption that the noise has a moment of order $2p$.

For moment of order larger than 2, we have the following result.

LEMMA 13. *Assume **A**1-**A**2-**A**3-**A**4(2p), for $p \geq 1$. There exist numerical constants $C_p, D_p \geq 2$ that only depend on p, such that, if $\gamma \in (0, 1/(LC_p))$, for all $k \in \mathbb{N}^*$ and $\theta_0 \in \mathbb{R}^d$*

$$\mathbb{E}^{1/p}\left[\left\|\theta_k^{(\gamma)} - \theta^*\right\|^{2p}\right] \leq (1 - 2\gamma\mu(1 - C_p\gamma L/2))^k \mathbb{E}^{1/p}\left[\|\theta_0 - \theta^*\|^{2p}\right] + \frac{D_p\gamma\tau_{2p}^2}{\mu} \; ,$$

where $(\theta_k^{(\gamma)})_{k\in\mathbb{N}}$ *is defined by* (1) *with initial condition* $\theta_0^{(\gamma)} = \theta_0$. *Moreover, the following bound holds*

$$\int_{\mathbb{R}^d} \|\theta - \theta^*\|^{2p} \, \pi_\gamma(\mathrm{d}\theta) \leq \left(D_p \gamma \tau_{2p}^2/\mu\right)^p \ . \tag{34}$$

REMARK 14.
- *Notably, Lemma 13 implies that* $\int_{\mathbb{R}^d} \|\theta - \theta^*\|^4 \, \pi_\gamma(\mathrm{d}\theta) = O(\gamma^2)$, *and thus* $\int_{\mathbb{R}^d} \|\theta - \theta^*\|^3 \, \pi_\gamma(\mathrm{d}\theta) = O(\gamma^{3/2})$. *We also note that* $\int_{\mathbb{R}^d} \|\theta - \theta^*\|^2 \pi_\gamma(\mathrm{d}\theta) = O(\gamma)$, *also implies by Jensen's inequality that* $\|\bar{\theta}_\gamma - \theta^*\|^2 = O(\gamma)$.
- *Note that there is no contradiction between* (34) *and Theorem 7, as for any* $p \geq 2$, *one has for* $g(\theta) = \|\theta - \theta^*\|^2$ *and* h_g *the solution to the Poisson equation, that* $h_g''(\theta^*) = 0$, *so that the first term in the development (of order* γ*) is indeed 0.*

PROOF. The proof is postponed to the supplementary document [17, Section S3]. □

LEMMA 15. *Let* $g : \mathbb{R}^d \to \mathbb{R}$ *satisfying* $\boldsymbol{A}6(1,p)$ *for* $p \in \mathbb{N}$. *Then for all* $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$|g(\theta_1) - g(\theta_2)| \leq a_g \|\theta_1 - \theta_2\| \left\{b_g + \|\theta_1 - \theta^*\|^p + \|\theta_2 - \theta^*\|^p\right\} \ .$$

PROOF. Let $\theta_1, \theta_2 \in \mathbb{R}^d$. By the mean value theorem, there exists $s \in [0,1]$ such that if $\eta_s = s\theta_1 + (1-s)\theta_2$ then

$$|g(\theta_1) - g(\theta_2)| = Dg(\eta_s)\left\{\theta_1 - \theta_2\right\} \ .$$

The proof is then concluded using $\mathbf{A}6(\ell, p)$ and

$$\|\eta_s - \theta^*\| \leq \max\left(\|\theta_1 - \theta^*\|, \|\theta_2 - \theta^*\|\right) \ .$$

□

PROPOSITION 16. *Let* $g : \mathbb{R}^d \to \mathbb{R}$ *satisfying* $\boldsymbol{A}6(1,p)$ *for* $p \in \mathbb{N}$. *Assume* $\boldsymbol{A}1$-$\boldsymbol{A}2$-$\boldsymbol{A}3$-$\boldsymbol{A}4(2p)$. *Let* $C_p \geq 2$ *be given by Lemma 13 and only depending on* p. *For all* $\gamma \in (0, 1/(LC_p))$, *for all initial point* $\theta_0 \in \mathbb{R}^d$, *there exists* C_g *independent of* θ_0 *such that for all* $k \geq 1$:

$$\left|\mathbb{E}\left[k^{-1}\sum_{i=1}^k \left\{g(\theta_i^{(\gamma)})\right\}\right] - \int_{\mathbb{R}^d} g(\theta)\pi_\gamma(\mathrm{d}\theta)\right| \leq C_g(1 + \|\theta_0 - \theta^*\|^p)/k \ .$$

PROOF. The proof is postponed to the supplementary document [17, Section S4].

□

6.3. *Proof of Proposition 3.*

PROOF OF PROPOSITION 3. By Lemma 9, we have $\int_{\mathbb{R}^d} f'(\theta)\pi_\gamma(\mathrm{d}\theta) = 0$. Since f' is linear, we get $f'(\bar{\theta}_\gamma) = 0$, which implies by **A**1 that $\bar{\theta}_\gamma = \theta^*$.

Let $\gamma \in (0, 2/L)$ and $(\theta_k^{(\gamma)})_{k\in\mathbb{N}}$ given by (1) with $\theta_0^{(\gamma)}$ distributed according to π_γ independent of $(\varepsilon_k)_{k\in\mathbb{N}^*}$. Note that if $f = f_\Sigma$, (1) implies for $k = 1$:

$$(\theta_1^{(\gamma)} - \theta^*)^{\otimes 2} = \left((\mathrm{Id} - \gamma\Sigma)\left(\theta_0^{(\gamma)} - \theta^*\right) + \gamma\varepsilon_1(\theta_0^{(\gamma)})\right)^{\otimes 2} .$$

Taking the expectation, using **A**3, $\theta_0^{(\gamma)}$ is independent of ε_1 and $\pi_\gamma R_\gamma = \pi_\gamma$, we get

$$\int_{\mathbb{R}^d} (\theta - \theta^*)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta) = (\mathrm{Id} - \gamma\Sigma)\left[\int_{\mathbb{R}^d} (\theta - \theta^*)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta)\right](\mathrm{Id} - \gamma\Sigma) + \gamma^2 \int_{\mathbb{R}^d} \mathcal{C}(\theta)\pi_\gamma(\mathrm{d}\theta) ,$$

$$(\Sigma\otimes\mathrm{Id} + \mathrm{Id}\otimes\Sigma - \gamma\Sigma\otimes\Sigma)\left[\int_{\mathbb{R}^d} (\theta - \theta^*)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta)\right] = \gamma\int_{\mathbb{R}^d} \mathcal{C}(\theta)\pi_\gamma(\mathrm{d}\theta) . \tag{35}$$

It remains to show that $(\Sigma\otimes\mathrm{Id} + \mathrm{Id}\otimes\Sigma - \gamma\Sigma\otimes\Sigma)$ is invertible. To show this result, we just claim that it is a symmetric positive definite operator. Indeed, since $\gamma < 2L^{-1}$, $\mathrm{Id} - (\gamma/2)\Sigma$ is symmetric positive definite and is diagonalizable with the same orthogonal vectors $(\mathbf{f}_i)_{i\in\{0,\ldots,d\}}$ as Σ. If we denote by $(\lambda_i)_{i\in\{0,\ldots,d\}}$, then we get that $(\Sigma\otimes\mathrm{Id} + \mathrm{Id}\otimes\Sigma - \gamma\Sigma\otimes\Sigma) = \Sigma\otimes(\mathrm{Id} - \gamma/2\Sigma) + (\mathrm{Id} - \gamma/2\Sigma)\otimes\Sigma$ is also diagonalizable in the orthogonal basis of $\mathbb{R}^d\otimes\mathbb{R}^d$, $(\mathbf{f}_i\otimes\mathbf{f}_j)_{i,j\in\{0,\ldots,d\}}$ and $(\lambda_i(1-\gamma\lambda_j) + \lambda_j(1-\gamma\lambda_i))_{i,j\in\{0,\ldots,d\}}$ are its eigenvalues. ☐

Note that in the case of the regression setting described in Example 1, we can specify Proposition 3 as follows.

PROPOSITION 17. *Assume that f is an objective function of a least-square regression problem, i.e. with the notations of Example 1, $f = f_\Sigma$, $\Sigma = \mathbb{E}[XX^\top]$ and ε_k are defined by (6). Assume* ***A****1-*__*A*__*2-*__*A*__*3-*__*A*__*4(4) and let r defined by (8). We have for all $\gamma \in \left(0, 1/r^2\right)$,*

$$(\Sigma\otimes\mathrm{Id} + \mathrm{Id}\otimes\Sigma - \gamma\mathbf{T})\left[\int_{\mathbb{R}^d} (\theta - \theta^*)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta)\right] = \gamma\mathbb{E}[\xi_1^{\otimes 2}] ,$$

where $\mathbf{T}$ and ξ_1 are defined by (18) and (7) respectively.

PROOF. The proof follows the same line as the proof of Proposition 3 and is omitted. ☐

6.4. *Proof of Theorem 4.* We preface the proof by a couple of preliminary lemmas.

LEMMA 18. *Assume* **A**1-**A**2-**A**3-**A**4$(6 \vee 2k_\varepsilon)$-**A**5 *and let* $\gamma \in (0, 2/L)$. *Then*

$$\bar{\theta}_\gamma - \theta^* = \gamma f''(\theta^*)^{-1} f'''(\theta^*) \mathbf{A} \left[\int_{\mathbb{R}^d} \{\mathcal{C}(\theta)\} \pi_\gamma(\mathrm{d}\theta) \right] + O(\gamma^{3/2}) , \tag{36}$$

where $\mathbf{A}$ *is defined by* (17), $\bar{\theta}_\gamma$ *and* $\mathcal{C}$ *are given by* (3) *and* (5) *respectively.*

PROOF. Let $\gamma \in (0, 2/L)$ and $(\theta_k^{(\gamma)})_{k\in\mathbb{N}}$ given by (1) with $\theta_0^{(\gamma)}$ distributed according to π_γ independent of $(\varepsilon_k)_{k\in\mathbb{N}^*}$. For conciseness, in the rest of the proof, we skip the explicit dependence in γ in $\theta_i^{(\gamma)}$: we only denote it θ_i.

First by a third order Taylor expansion with integral remainder of f' around θ^*, we have that for all $x \in \mathbb{R}^d$,

$$f'(\theta) = f''(\theta^*)(\theta - \theta^*) + (1/2) f'''(\theta^*)(\theta - \theta^*)^{\otimes 2} + \mathcal{R}_1(\theta) , \tag{37}$$

where $\mathcal{R}_1 : \mathbb{R}^d \to \mathbb{R}^d$ satisfies

$$\sup_{\theta \in \mathbb{R}^d} \{ \|\mathcal{R}_1(\theta)\| / \|\theta - \theta^*\|^3 \} < +\infty . \tag{38}$$

It follows from Lemma 9, taking the integral with respect to π_γ,

$$0 = \int_{\mathbb{R}^d} \left\{ f''(\theta^*)(\theta - \theta^*) + (1/2) f'''(\theta^*)(\theta - \theta^*)^{\otimes 2} + \mathcal{R}_1(\theta) \right\} \pi_\gamma(\mathrm{d}\theta) .$$

Using (38), Lemma 13 and Hölder inequality, we get

$$f''(\theta^*)(\bar{\theta}_\gamma - \theta^*) + (1/2) f'''(\theta^*) \left[\int_{\mathbb{R}^d} (\theta - \theta^*)^{\otimes 2} \pi_\gamma(\mathrm{d}\theta) \right] = O(\gamma^{3/2}) . \tag{39}$$

Moreover, we have by a second order Taylor expansion with integral remainder of f' around θ^*,

$$\theta_1 - \theta^* = \theta_0 - \theta^* - \gamma \big[f''(\theta^*)(\theta_0 - \theta^*) + \varepsilon_1(\theta_0) + \mathcal{R}_2(\theta_0) \big] ,$$

where $\mathcal{R}_2 : \mathbb{R}^d \to \mathbb{R}^d$ satisfies

$$\sup_{\theta \in \mathbb{R}^d} \{ \|\mathcal{R}_2(\theta)\| / \|\theta - \theta^*\|^2 \} < +\infty . \tag{40}$$

Taking the second order moment of this equation, and using **A**3, θ_0 is independent of ε_1, (40), Lemma 13 and Hölder inequality, we get

$$\int_{\mathbb{R}^d}(\theta-\theta^*)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta) = (\mathrm{Id} - \gamma f''(\theta^*))\left[\int_{\mathbb{R}^d}(\theta-\theta^*)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta)\right](\mathrm{Id} - \gamma f''(\theta^*)) + \gamma^2\int_{\mathbb{R}^d}\mathcal{C}(\theta)\pi_\gamma(\mathrm{d}\theta) + O(\gamma^{5/2}).$$

This leads to:

$$\int_{\mathbb{R}^d}(\theta-\theta^*)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta) = \gamma\mathbf{A}\left[\int_{\mathbb{R}^d}\mathcal{C}(\theta)\pi_\gamma(\mathrm{d}\theta)\right] + O(\gamma^{3/2})\ .$$

Combining this result and (39), we have that (36) holds if the operator $(f''(\theta^*)\otimes\mathrm{Id} + \mathrm{Id}\otimes f''(\theta^*) - \gamma f''(\theta^*)\otimes f''(\theta^*))$ is invertible. To show this result, like in the quadratic case, we just claim that it is a symmetric positive definite operator. Indeed, since $\gamma < 2L^{-1}$, by **A**1, $\mathrm{Id} - (\gamma/2)f''(\theta^*)$ is symmetric positive definite and is diagonalizable with the same orthogonal vectors $(\mathbf{f}_i)_{i\in\{0,\ldots,d\}}$ as $f''(\theta^*)$. If we denote by $(\lambda_i)_{i\in\{0,\ldots,d\}}$, then we get that $(f''(\theta^*)\otimes\mathrm{Id} + \mathrm{Id}\otimes f''(\theta^*) - \gamma f''(\theta^*)\otimes f''(\theta^*)) = f''(\theta^*)\otimes(\mathrm{Id} - \gamma/2 f''(\theta^*)) + (\mathrm{Id} - \gamma/2 f''(\theta^*))\otimes f''(\theta^*)$ is also diagonalizable in the orthogonal basis of $\mathbb{R}^d\otimes\mathbb{R}^d$, $(\mathbf{f}_i\otimes\mathbf{f}_j)_{i,j\in\{0,\ldots,d\}}$ and $(\lambda_i(1-\gamma\lambda_j) + \lambda_j(1-\gamma\lambda_i))_{i,j\in\{0,\ldots,d\}}$ are its eigenvalues. □

LEMMA 19. *Assume* ***A**1-**A**2-**A**3-**A**4$(6\vee[2(k_\varepsilon+1)])$-**A**5. It holds as $\gamma\to 0$,*

$$\int_{\mathbb{R}^d}\mathcal{C}(\theta)\pi_\gamma(\mathrm{d}\theta) = \mathcal{C}(\theta^*) + O(\gamma)\ ,$$
$$\int_{\mathbb{R}^d}\mathcal{C}(\theta)\otimes\{\theta-\theta^*\}\pi_\gamma(\mathrm{d}\theta) = \mathcal{C}(\theta^*)\{\bar{\theta}_\gamma - \theta^*\} + O(\gamma)\ ,$$

where $\mathcal{C}$ is given by (5).

PROOF. By a second order Taylor expansion around θ^* of $\mathcal{C}$ and using **A**5, we get for all $x\in\mathbb{R}^d$ that

$$\mathcal{C}(x) - \mathcal{C}(\theta^*) = \mathcal{C}'(\theta^*)\{x-\theta^*\} + \mathcal{R}_1(x)\ ,$$

where $\mathcal{R}_1 : \mathbb{R}^d\to\mathbb{R}^d$ satisfies $\sup_{x\in\mathbb{R}^d}\|\mathcal{R}_1(x)\|/(\|x-\theta^*\|^2 + \|x+\theta^*\|^{k_\varepsilon+2}) < +\infty$. Taking the integral with respect to π_γ and using Lemma 18-Lemma 13 concludes the proof. □

Proof of Theorem 4. Let $\gamma \in (0, 2/L)$ and $(\theta_k^{(\gamma)})_{k\in\mathbb{N}}$ given by (1) with $\theta_0^{(\gamma)}$ distributed according to π_γ independent of $(\varepsilon_k)_{k\in\mathbb{N}^*}$. For conciseness, in the rest of the proof, we skip the explicit dependence in γ in $\theta_i^{(\gamma)}$: we only denote it θ_i.

The proof consists in showing that the residual term in (36) of Lemma 18 is of order $O(\gamma^2)$ and not only $O(\gamma^{3/2})$. Note that we have already prove that $\bar{\theta}_\gamma - \theta^* = O(\gamma)$. To find the next term in the development, we develop further each of the terms. By a fourth order Taylor expansion with integral remainder of f' around θ^*, and using **A**2, we have

$$\begin{aligned}\theta_1 - \theta^* =& \theta_0 - \theta^* - \gamma\big[f''(\theta^*)(\theta_0 - \theta^*) + (1/2)f^{(3)}(\theta^*)(\theta_0 - \theta^*)^{\otimes 2} \\ & + (1/6)f^{(4)}(\theta^*)(\theta_0 - \theta^*)^{\otimes 3} + \varepsilon_1(\theta_0) + \mathcal{R}_3(\theta)\big] \,,\end{aligned} \tag{41}$$

where $\mathcal{R}_3 : \mathbb{R}^d \to \mathbb{R}^d$ satisfies $\sup_{x\in\mathbb{R}^d} \|\mathcal{R}_3(x)\| / \|x - \theta^*\|^4 < +\infty$. Therefore taking the expectation and using **A**3-Lemma 13 we get

$$\begin{aligned}f''(\theta^*)(\bar{\theta}_\gamma - \theta^*) = & -(1/2)f^{(3)}(\theta^*)\int_{\mathbb{R}^d}(\theta - \theta^*)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta) \\ & - (1/6)f^{(4)}(\theta^*)\int_{\mathbb{R}^d}(\theta - \theta^*)^{\otimes 3}\pi_\gamma(\mathrm{d}\theta) + O(\gamma^2) \,.\end{aligned} \tag{42}$$

Since $f''(\theta^*)$ is invertible by **A**1, To get the next term in the development, we show that

(a) $\int_{\mathbb{R}^d}(\theta - \theta^*)^{\otimes 3}\pi_\gamma(\mathrm{d}\theta) = \blacksquare\gamma^2 + o(\gamma^2)$.
(b) $\int_{\mathbb{R}^d}(\theta - \theta^*)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta) = \square\gamma + \triangle\gamma^2 + o(\gamma^2)$, for $\square$ given in (16), proving (16).

(a) Denote for $i = 0, 1$, $\eta_i = \theta_i - \theta^*$. By (37)-(38), Lemma 13 and **A**3-**A**4(12), we get

$$\begin{aligned}\mathbb{E}[\eta_1^{\otimes 3}] &= \mathbb{E}\left[\left\{(\mathrm{Id} - \gamma f''(\theta^*))\eta_0 - \gamma\varepsilon_1(\theta_0) - \gamma f'''(\theta^*)\eta_0^{\otimes 2} + \mathcal{R}_1(\theta_0)\right\}^{\otimes 3}\right] \\ &= \mathbb{E}\big[\{(\mathrm{Id} - \gamma f''(\theta^*))\eta_0\}^{\otimes 3} + \gamma^2\{\varepsilon_1(\theta_0)\}^{\otimes 2} \otimes \{(\mathrm{Id} - \gamma f''(\theta^*))\eta_0\} \\ &+ \gamma\{(\mathrm{Id} - \gamma f''(\theta^*))\eta_0\}^{\otimes 2} \otimes \{f'''(\theta^*)\eta_0^{\otimes 2}\} \\ &\quad + \gamma\{f'''(\theta^*)\eta_0^{\otimes 2}\} \otimes \{(\mathrm{Id} - \gamma f''(\theta^*))\eta_0\}^{\otimes 2}\big] + O(\gamma^3) \\ &= \mathbb{E}\big[\{(\mathrm{Id} - \gamma f''(\theta^*))\eta_0\}^{\otimes 3} + \gamma^2\{\varepsilon_1(\theta_0)\}^{\otimes 2} \otimes \{(\mathrm{Id} - \gamma f''(\theta^*))\eta_0\}\big] + O(\gamma^3) \\ &= \mathbb{E}\big[\{\eta_0\}^{\otimes 3}\big] + \mathbb{E}\big[\gamma\mathbf{B}\{\eta_0\}^{\otimes 3} + \gamma^2\{\varepsilon_1(\theta_0)\}^{\otimes 2} \otimes \{(\mathrm{Id} - \gamma f''(\theta^*))\eta_0\}\big] + O(\gamma^3) \,,\end{aligned}$$

where $\mathbf{B} \in \mathrm{L}(\mathbb{R}^{d^3}, \mathbb{R}^{d^3})$ is defined by

$$\mathbf{B} = f''(\theta^*) \otimes \mathrm{Id} \otimes \mathrm{Id} + \mathrm{Id} \otimes f''(\theta^*) \otimes \mathrm{Id} + \mathrm{Id} \otimes \mathrm{Id} \otimes f''(\theta^*) \,.$$

Using **A**1 and the same reasoning as to show that **A** in (17), is well defined, we get that **B** is invertible. Then since η_0 and η_1 has the same distribution π_γ, we get

$$\int_{\mathbb{R}^d} (\theta - \theta^*)^{\otimes 3} \pi_\gamma(\mathrm{d}\theta)$$
$$= \gamma \mathbf{B}^{-1} \left[\int_{\mathbb{R}^d} \{\mathcal{C}(\theta)\} \otimes \{(\mathrm{Id} - \gamma f''(\theta^*))(\theta - \theta^*)\} \pi_\gamma(\mathrm{d}\theta) \right] + O(\gamma^2) \,.$$

By Lemma 19, we get

$$\int_{\mathbb{R}^d} (\theta - \theta^*)^{\otimes 3} \pi_\gamma(\mathrm{d}\theta) = \gamma \mathbf{B}^{-1} \left[\{\mathcal{C}(\theta^*)\} \otimes \{(\mathrm{Id} - \gamma f''(\theta^*))(\bar{\theta}_\gamma - \theta^*)\} \right] + O(\gamma^2) \,.$$

Combining this result and (36) implies (a).

(b) First, we have using (41), **A**3 and Lemma 13 that:

$$\mathbb{E}[(\theta_1 - \theta^*)^{\otimes 2}] = \mathbb{E}\left[(\theta_0 - \theta^*)^{\otimes 2} - \gamma(\mathrm{Id} \otimes f''(\theta^*) + f''(\theta^*) \otimes \mathrm{Id})(\theta - \theta^*)^{\otimes 2} \right.$$
$$+ (\gamma/2)(\theta_0 - \theta^*) \otimes \{f^{(3)}(\theta^*)(\theta_0 - \theta^*)^{\otimes 2}\}$$
$$\left. + (\gamma/2)\{f^{(3)}(\theta^*)(\theta_0 - \theta^*)^{\otimes 2}\} \otimes (\theta_0 - \theta^*) + \gamma^2 \varepsilon_1(\theta_0)^{\otimes 2}(\theta_0) \right] + O(\gamma^3) \,.$$

Since θ_0 and θ_1 follow the same distribution π_γ, it follows that

$$\gamma(\mathrm{Id} \otimes f''(\theta^*) + f''(\theta^*) \otimes \mathrm{Id}) \left[\int_{\mathbb{R}^d} (\theta - \theta^*)^{\otimes 2} \pi_\gamma(\mathrm{d}\theta) \right]$$
$$= O(\gamma^3) + \int_{\mathbb{R}^d} \left[(\gamma/2)(\theta - \theta^*) \otimes \{f^{(3)}(\theta^*)(\theta - \theta^*)^{\otimes 2}\} \right.$$
$$\left. + \frac{\gamma}{2}\{f^{(3)}(\theta^*)(\theta - \theta^*)^{\otimes 2}\} \otimes (\theta - \theta^*) + \gamma^2 \varepsilon_1(\theta_0)^{\otimes 2}(\theta_0) \right] \pi_\gamma(\mathrm{d}\theta) \,. \tag{43}$$

Then by linearity of $f'''(\theta^*)$ and using (a) we get (b).

Finally the proof of (15) follows from combining the results of (a)-(b) in (42).

□

6.5. *Proof of Theorem 5.* Theorem 5 follows from the following more general result taking $\varphi : \theta \mapsto \theta - \theta^*$.

THEOREM 20. *Let $\varphi : \mathbb{R}^d \to \mathbb{R}^q$ be a Lipschitz continuous function. Assume **A**1-**A**2-**A**3-**A**4(4) and let $\gamma \in (0, 1/(2L))$. Then setting $\rho = (1 - 2\mu\gamma(1 - \gamma L))^{1/2}$, for any starting point $\theta_0 \in \mathbb{R}^d$, $k \in \mathbb{N}^*$*

$$\mathbb{E}\left[k^{-1} \sum_{i=0}^{k-1} \varphi(\theta_i^{(\gamma)}) \right] = \pi_\gamma(\varphi) + (1/k)\psi_\gamma(\theta_0) + O(k^{-2}) \,,$$

and if $\pi_\gamma(\varphi) = 0$*,*

$$\mathbb{E}\left[\left\{k^{-1}\sum_{i=0}^{k-1}\varphi(\theta_i^{(\gamma)})\right\}^{\otimes 2}\right] = \frac{1}{k}\pi_\gamma\left[\psi_\gamma^{\otimes 2} - (\psi_\gamma - \varphi)^{\otimes 2}\right] \\ - \frac{1}{k^2}\left[\pi_\gamma(\varpi_\gamma\varphi^\top + \varphi\varpi_\gamma^\top) + \chi_\gamma^2(\theta_0) - \chi_\gamma^1(\theta_0)\right] + O(k^{-3}) ,$$

where ψ_γ*,* ϖ_γ*,* χ_γ^1*,* χ_γ^2 *are solutions of the Poisson equation* (26) *associated with* φ*,* ψ_γ*,* $\psi_\gamma^{\otimes 2}$ *and* $(\psi_\gamma - \varphi)^{\otimes 2}$ *respectively.*

PROOF. In the proof C will denote generic constants which can change from line to line. In addition, we skip the dependence on γ for $\theta_k^{(\gamma)}$, simply denoted θ_k. Let $\theta_0 \in \mathbb{R}^d$. By Lemma 8, ψ_γ exists and is Lipschitz continuous, and using Proposition 2-(b), $\pi_\gamma(\psi_\gamma) = 0$, we have that $R_\gamma^k\psi_\gamma(\theta_0) = O(\rho^k)$, with $\rho := (1 - 2\mu\gamma(1-\gamma L))^{1/2}$. Therefore, setting $\Phi_k = k^{-1}\sum_{i=0}^{k-1}\varphi(\theta_i)$,

$$\begin{aligned}\mathbb{E}[\Phi_k] &= k^{-1}\sum_{i=0}^{k-1}\mathbb{E}\left[\varphi(\theta_i)\right] = k^{-1}\sum_{i=0}^{k-1}R_\gamma^i\varphi(\theta_0) \\ &= \pi_\gamma(\varphi) + k^{-1}\sum_{i=0}^{k-1}(R_\gamma^i\varphi(\theta_0) - \pi_\gamma(\varphi))) \\ &= \pi_\gamma(\varphi) + k^{-1}\psi_\gamma(\theta_0) - R_\gamma^k\psi_\gamma(\theta_0) = \pi_\gamma(\varphi) + k^{-1}\psi_\gamma(\theta_0) + O(\rho^k) .\end{aligned}$$

We now consider the Poisson solution associated with $\varphi\varphi^\top$, χ_γ^3. By Lemma 11, such a function exists and satisfies $\pi_\gamma(\chi_\gamma^3) = 0$, $R_\gamma^k\chi_\gamma^3(\theta_0) = O(\rho^k)$. Therefore, we obtain using in addition the Markov property:

$$\begin{aligned}&\mathbb{E}[\Phi_k\Phi_k^\top] = \frac{1}{k^2}\sum_{i,j=0}^{k-1}\mathbb{E}\left[\varphi(\theta_i)\varphi(\theta_j)^\top\right] \\ &= \frac{1}{k^2}\sum_{i=0}^{k-1}\left(\mathbb{E}\left[\varphi(\theta_i)\varphi(\theta_i)^\top\right] + \sum_{j=i+1}^{k-1}\left\{\mathbb{E}\left[\varphi(\theta_i)\varphi(\theta_j)^\top\right] + \mathbb{E}\left[\varphi(\theta_j)\varphi(\theta_i)^\top\right]\right\}\right) \\ &= -\frac{1}{k^2}\sum_{i=0}^{k-1}R_\gamma^i(\varphi\varphi^\top)(\theta_0) \\ &\quad + \frac{1}{k^2}\sum_{i=0}^{k-1}\left(\sum_{j=i+1}^{k-1}\left\{\mathbb{E}\left[\varphi(\theta_i)\varphi(\theta_j)^\top\right] + \mathbb{E}\left[\varphi(\theta_j)\varphi(\theta_i)^\top\right]\right\}\right)\end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{k}\pi_\gamma(\varphi\varphi^\top) - \frac{1}{k^2}\sum_{i=0}^{\infty}\left\{R_\gamma^i(\varphi\varphi^\top)(\theta_0) - \pi_\gamma(\varphi\varphi^\top)\right\} + O(\rho^k) \\
&\quad + \frac{1}{k^2}\sum_{i=0}^{k-1}\left(\sum_{j=i+1}^{k-1}\left\{\mathbb{E}\left[\varphi(\theta_i)\varphi(\theta_j)^\top\right] + \mathbb{E}\left[\varphi(\theta_j)\varphi(\theta_i)^\top\right]\right\}\right) \\
&= -\frac{1}{k}\pi_\gamma(\varphi\varphi^\top) - \frac{1}{k^2}\chi_\gamma^3(\theta_0) + O(\rho^k) \\
&+ \frac{1}{k^2}\sum_{i=0}^{k-1}\left(\sum_{j=0}^{k-1-i}\left\{\mathbb{E}\left[\varphi(\theta_i)(R_\gamma^j\varphi(\theta_i))^\top\right] + \mathbb{E}\left[R_\gamma^j\varphi(\theta_i)\varphi(\theta_i)^\top\right]\right\}\right) .
\end{aligned}
$$

Thus using that for all $N \in \mathbb{N}$ and $\theta \in \mathbb{R}^d$, $\sum_{j=0}^N R_\gamma^j\varphi(\theta) = \sum_{j=0}^N\{R_\gamma^j\psi_\gamma(\theta) - R_\gamma^{j+1}\psi_\gamma(\theta)\} = \psi_\gamma(\theta) - R_\gamma^{N+1}\psi_\gamma(\theta)$, we get

$$
\begin{aligned}
\mathbb{E}[\Phi_k\Phi_k^\top] &= -\frac{1}{k}\pi_\gamma(\varphi\varphi^\top) - \frac{1}{k^2}\chi_\gamma^3(\theta_0) \\
&\quad + \frac{1}{k^2}\sum_{i=0}^{k-1}\left\{R_\gamma^i\left[\varphi\psi_\gamma^\top - \varphi(R_\gamma^{k-i}\psi_\gamma)^\top\right](\theta_0)\right\} \\
&\quad + \frac{1}{k^2}\sum_{i=0}^{k-1}\left\{R_\gamma^i\left[\psi_\gamma\varphi^\top - R_\gamma^{k-i}\psi_\gamma\varphi^\top\right](\theta_0)\right\} + O(\rho^k) . \qquad (44)
\end{aligned}
$$

Moreover, since φ is Lipschitz continuous and $R_\gamma^N\psi_\gamma$ is $C\rho^N$-Lipschitz continuous and we have $\sup_{x\in\mathbb{R}^d}\{R_\gamma^N\psi_\gamma(x)/\|x\|\} \leq C\rho^N$ by Lemma 8, we get for all $x, y \in \mathbb{R}^d$ and $N \in \mathbb{N}$,

$$
\left\|\varphi(R_\gamma^N\psi_\gamma)^\top(x) - \varphi(R_\gamma^N\psi_\gamma)^\top(y)\right\| \leq C\rho^N\|x-y\|(1+\|x\|+\|y\|) . \qquad (45)
$$

Then, we obtain by Lemma 11

$$
\begin{aligned}
\frac{1}{k}\sum_{i=0}^{k-1}R_\gamma^i[\varphi(R_\gamma^{k-i}\psi_\gamma)^\top](\theta_0) &= \frac{1}{k}\sum_{i=0}^{k-1}[R_\gamma^i - \pi_\gamma][\varphi(R_\gamma^{k-i}\psi_\gamma)^\top](\theta_0) \\
&\quad \frac{1}{k}\sum_{i=0}^{k-1}\pi_\gamma[\varphi(R_\gamma^{k-1}\psi_\gamma)^\top](\theta_0) \\
&= (C/k)(1+\|\theta_0\|)\sum_{i=0}^{k-1}\rho^k + \pi_\gamma(\varphi\varpi_\gamma^\top)/k + O(k^{-2}) , \qquad (46)
\end{aligned}
$$

using $\pi_\gamma(\psi_\gamma) = 0$, $\sum_{i=0}^{+\infty}R_\gamma^i\psi_\gamma(\theta) = \varpi_\gamma(\theta)$, for all $\theta \in \mathbb{R}^d$, where ϖ_γ is the

Poisson solution associated with ψ_γ. Similarly, we have

$$\frac{1}{k}\sum_{i=0}^{k-1} R_\gamma^i[R_\gamma^{k-i}\psi_\gamma\varphi^\top](\theta_0) = \pi_\gamma(\varpi_\gamma\varphi^\top)/k + O(k^{-2})$$

$$\frac{1}{k}\sum_{i=0}^{k-1}\left\{R_\gamma^i[\varphi\psi_\gamma^\top](\theta_0) - \pi_\gamma[\varphi\psi_\gamma^\top]\right\} = \chi_\gamma^4(\theta_0) + O(k^{-2}) \tag{47}$$

$$\frac{1}{k}\sum_{i=0}^{k-1}\left\{R_\gamma^i[\psi_\gamma\varphi^\top](\theta_0) - \pi_\gamma[\psi_\gamma\varphi^\top]\right\} = \chi_\gamma^5(\theta_0) + O(k^{-2}) \ ,$$

where χ_γ^4 and χ_γ^5 are the Poisson solution associated with $\varphi\psi_\gamma^\top$ and $\psi_\gamma\varphi^\top$ respectively. Combining (46)-(47) in (44), we obtain

$$\begin{aligned}\mathbb{E}[\Phi_k\Phi_k^\top] &= \frac{1}{k}[\pi_\gamma(\varphi\psi_\gamma^\top) + \pi_\gamma(\psi_\gamma\varphi^\top) - \pi_\gamma(\varphi\varphi^\top)] + O(k^{-3}) \\ &\quad - \frac{1}{k^2}[\pi_\gamma(\varphi\varpi_\gamma^\top) + \pi_\gamma(\varpi_\gamma\varphi^\top) + \chi_\gamma^3(\theta_0) - \chi_\gamma^4(\theta_0) - \chi_\gamma^5(\theta_0)] \ .\end{aligned} \tag{48}$$

First note that

$$-\varphi\varphi^\top + \varphi\psi_\gamma^\top + \psi_\gamma\varphi^\top = -(\varphi-\psi_\gamma)(\varphi-\psi_\gamma)^\top + \psi_\gamma\psi_\gamma^\top \ . \tag{49}$$

In addition, by Lemma 11 and definition, we have for all θ_0

$$\begin{aligned}&\chi_\gamma^3(\theta_0) - \chi_\gamma^4(\theta_0) - \chi_\gamma^5(\theta_0) \\ &= \sum_{i=1}^{+\infty}\left\{R_\gamma^i[\varphi\varphi^\top - \varphi\psi_\gamma^\top - \psi_\gamma\varphi^\top](\theta_0) - \pi_\gamma[\varphi\varphi^\top - \varphi\psi_\gamma^\top - \psi_\gamma\varphi^\top]\right\} \\ &= \sum_{i=1}^{+\infty}\left\{R_\gamma^i[(\varphi-\psi_\gamma)(\varphi-\psi_\gamma)^\top - \psi_\gamma\psi_\gamma^\top](\theta_0) - \pi_\gamma[(\varphi-\psi_\gamma)(\varphi-\psi_\gamma)^\top - \psi_\gamma\psi_\gamma^\top]\right\} \\ &= \chi^2(\theta_0) - \chi^1(\theta_0) \ .\end{aligned}$$

Combining this result and (49) in (48) concludes the proof. □

6.6. *Proof of Theorem 7.* Before giving the proof of Theorem 7, we need several results regarding Poisson solutions associated with the gradient flow ODE (20).

6.6.1. *Regularity of the gradient flow and estimates on Poisson solution.* Let $\ell \in \mathbb{N}^*$ and consider the following assumption.

A9 (ℓ). *$f \in C^{\ell}(\mathbb{R}^d)$ and there exists $M \geq 0$ such that for all $i \in \{2, \dots, \ell\}$, $\sup_{\theta \in \mathbb{R}^d} \left\| f^{(i)}(\theta) \right\| \leq \bar{L}$.*

LEMMA 21. *Assume **A**1 and **A**9($\ell + 1$) for $\ell \in \mathbb{N}^*$.*

a) For all $t \geq 0$, $\varphi_t \in \mathrm{C}^{\ell}(\mathbb{R}^d, \mathbb{R}^d)$, where $(\varphi_t)_{t \in \mathbb{R}_+}$ is the differential flow associated with (19). In addition, for all $\theta \in \mathbb{R}$, $t \mapsto \varphi_t^{(\ell)}(\theta)$ satisfies the following ordinary differential equation,

$$\left.\frac{\mathrm{d}\varphi_s^{(\ell)}(\theta)}{\mathrm{d}s}\right|_{s=t} = D^{\ell} \left\{ f' \circ \varphi_t \right\}(\theta) \ , \ \text{for all } t \geq 0 \ ,$$

with $\varphi_0^{'} = \mathrm{Id}$ and $\varphi_0^{(\ell)} = 0$ for $\ell \geq 2$.

b) For all $t \geq 0$ and $\theta \in \mathbb{R}^d$, $\|\varphi_t(\theta) - \theta^*\|^2 \leq \mathrm{e}^{-2\mu t} \|\theta - \theta^*\|^2$.

c) If $\ell \geq 2$, for all $t \geq 0$,

$$\varphi_t'(\theta^*) = \mathrm{e}^{-f''(\theta^*)t} \ .$$

d) If $\ell \geq 3$, for all $t \geq 0$ and $i, j, l \in \{1, \dots, d\}$,

$$\begin{aligned} &\left\langle {\varphi_t}''(\theta^*) \left\{ \mathbf{f}_i \otimes \mathbf{f}_j \right\}, \mathbf{f}_l \right\rangle \\ &\qquad = \begin{cases} \frac{\mathrm{e}^{-\lambda_l t} - \mathrm{e}^{-(\lambda_i + \lambda_j)t}}{\lambda_l - \lambda_i - \lambda_j} f^{(3)}(\theta^*) \left\{ \mathbf{f}_i \otimes \mathbf{f}_j \otimes \mathbf{f}_l \right\} & \text{if } \lambda_l \neq \lambda_i + \lambda_j \\ -t\mathrm{e}^{-\lambda_l t} f^{(3)}(\theta^*) \left\{ \mathbf{f}_i \otimes \mathbf{f}_j \otimes \mathbf{f}_l \right\} & \text{otherwise} \ , \end{cases} \end{aligned}$$

where $\{\mathbf{f}_1, \dots, \mathbf{f}_d\}$ and $\{\lambda_1, \dots, \lambda_d\}$ are the eigenvectors and the eigenvalues of $f''(\theta^*)$ respectively satisfying for all $i \in \{1, \dots, d\}$, $f''(\theta^*)\mathbf{f}_i = \lambda_i \mathbf{f}_i$.

PROOF. a) This is a fundamental result on the regularity of flows of autonomous differential equations, see, *e.g.* [25, Theorem 4.1 Chapter V]

b) Let $\theta \in \mathbb{R}^d$. Differentiate $\|\varphi_t(\theta)\|^2$ with respect to t and using **A**1, that f is at least continuously differentiable and Grönwall's inequality concludes the proof.

c) By a) and since θ^* is an equilibrium point, for all $x \in \mathbb{R}^d$, $\xi_t^x(\theta^*) = \varphi_t'(\theta^*)\{x\}$ satisfies the following ordinary differential equation

$$\dot{\xi_s^x}(\theta^*) = -f''(\varphi_s(\theta^*))\xi_s^x(\theta^*)\mathrm{d}s = -f''(\theta^*)\xi_s^x(\theta^*)\mathrm{d}s \ . \tag{50}$$

with $\xi_0^x(\theta^*) = x$. The proof then follows from uniqueness of the solution of (50).

d) By a), for all $x_1, x_2 \in \mathbb{R}^d$, $\xi_t^{x_1,x_2}(\theta^*) = {\varphi_t}''(\theta^*)\left\{x_1 \otimes x_2\right\}$ satisfies the ordinary stochastic differential equation:

$$\frac{\mathrm{d}\xi_s^{x_1,x_2}}{\mathrm{d}s}(\theta^*) = -f^{(3)}(\varphi_s(\theta^*))\left\{{\varphi_s}'(\theta^*)x_1 \otimes {\varphi_s}'(\theta^*)x_2 \otimes \mathbf{e}_i\right\} \\ - f''(\varphi_s(\theta^*))\left\{\xi_s^{x_1,x_2} \otimes \mathbf{e}_i\right\} .$$

By c) and since θ^* is an equilibrium point we get that $\xi_t^{x_1,x_2}(\theta^*)$ satisfies

$$\frac{\mathrm{d}\xi_s^{x_1,x_2}}{\mathrm{d}s}(\theta^*) = -f^{(3)}(\theta^*)\left\{\mathrm{e}^{-f''(\theta^*)t}x_1 \otimes \mathrm{e}^{-f''(\theta^*)t}x_2 \otimes \mathbf{e}_i\right\} - f''(\theta^*)\left\{\xi_s^{x_1,x_2} \otimes \mathbf{e}_i\right\} .$$

Therefore we get for all $i, j, l \in \{1, \ldots, d\}$,

$$\frac{\mathrm{d}\left\langle \xi_s^{\mathbf{f}_i,\mathbf{f}_j}, \mathbf{f}_l\right\rangle}{\mathrm{d}s} = -f^{(3)}(\theta^*)\left\{\mathrm{e}^{-\lambda_i t}\mathbf{f}_i \otimes \mathrm{e}^{-\lambda_j t}\mathbf{f}_j \otimes \mathbf{f}_l\right\} - \lambda_l \left\langle \xi_s^{\mathbf{f}_i,\mathbf{f}_j}, \mathbf{f}_l\right\rangle .$$

This ordinary differential equation can be solved analytically which finishes the proof.

□

Under **A**1 and **A**9(ℓ), for any function $g : \mathbb{R}^d \to \mathbb{R}^q$, locally Lipschitz continuous, denote by h_g the solution of the continuous Poisson equation defined for all $\theta \in \mathbb{R}^d$ by

$$h_g(\theta) = \int_0^{\infty} (g(\varphi_s(\theta)) - g(\theta^*))dt . \tag{51}$$

Note that h_g is well-defined by Lemma 21-b) and since g is assumed to be locally-Lipschitz. In addition by (20), h_g satisfies

$$\mathcal{A}h_g(\theta) = g(\theta) - g(\theta^*) . \tag{52}$$

Define $h_{\mathrm{Id}} : \mathbb{R}^d \to \mathbb{R}^d$ for all $x \in \mathbb{R}^d$ by

$$h_{\mathrm{Id}}(\theta) = \int_0^{\infty} \left\{\varphi_s(\theta) - \theta^*\right\} dt . \tag{53}$$

Note that h_{Id} is also well-defined by Lemma 21-b).

LEMMA 22. *Let $g : \mathbb{R}^d \to \mathbb{R}$ satisfying **A**6(ℓ, p) for $\ell, p \in \mathbb{N}$, $\ell \geq 1$. Assume **A**1 and **A**9($\ell + 1$).*

a) Then for all $\theta \in \mathbb{R}^d$,

$$|h_g|(\theta) \leq a_g\left\{(b_g/\mu)\|\theta - \theta^*\| + (p\mu)^{-1}\|\theta - \theta^*\|^p\right\} .$$

b) *If $\ell \geq 2$, then $\nabla h_{\mathrm{Id}}(\theta^*) = (f''(\theta^*))^{-1}$. If $\ell \geq 3$, then for all $i, j \in \{1, \ldots, d\}$,*

$$\frac{\partial^2 h_{\mathrm{Id}}}{\partial\theta_i \partial\theta_j}(\theta^*) = \sum_{l=1}^{d} \Big[-f^{(3)}(\theta^*) \Big\{ \Big[\big(f''(\theta^*) \otimes \mathrm{Id} + \mathrm{Id} \otimes f''(\theta^*)\big)^{-1} \{\mathbf{e}_i \otimes \mathbf{e}_j\} \Big] \otimes \mathbf{e}_i \Big\} \times (f''(\theta^*))^{-1} \mathbf{e}_l \Big] .$$

PROOF. a) For all $\theta \in \mathbb{R}^d$, we have using Lemma 15 and (51)

$$|h_g(\theta)| \leq a_g \int_0^{+\infty} \|\varphi_s(\theta) - \theta^*\| \{b_g + \|\varphi_s(\theta) - \theta^*\|^p\} \, \mathrm{d}s .$$

The proof then follows from Lemma 21-b).

b) The proof is a direct consequence of Lemma 21-c)-d) and (51). □

THEOREM 23. *Let $g : \mathbb{R}^d \to \mathbb{R}$ satisfying $\boldsymbol{A}6(\ell, p)$ for $\ell, p \in \mathbb{N}$, $\ell \geq 2$. Assume $\boldsymbol{A}1$-$\boldsymbol{A}9(\ell + 1)$.*

a) *For all $i \in \{1, \ldots, \ell\}$, there exists $C_i \geq 0$ such that for all $\theta \in \mathbb{R}^d$ and $t \geq 0$,*

$$\left\| \varphi_t^{(i)}(\theta) \right\| \leq C_i \mathrm{e}^{-\mu t} .$$

b) *Furthermore, $h_g \in \mathrm{C}^\ell(\mathbb{R}^d)$ and for all $i \in \{0, \ldots, \ell\}$, there exists $C_i \geq 0$ such that for all $\theta \in \mathbb{R}^d$,*

$$\left\| h_g^{(i)}(\theta) \right\| \leq C_i \{1 + \|\theta - \theta^*\|^p\} .$$

PROOF. a) The proof is by induction on ℓ. By Lemma 21-a), for all $x \in \mathbb{R}^d$, and $\theta \in \mathbb{R}^d$, $\xi_t^x(\theta) = D\varphi_t(\theta)\{x\}$ satisfies

$$\frac{\mathrm{d}\xi_s^x(\theta)}{\mathrm{d}s}\Big|_{s=t} = -f''(\varphi_t(\theta))\xi_t^x(\theta) . \tag{54}$$

with $\xi_0^x(\theta) = x$. Now differentiating $s \to \|\xi_s^x(\theta)\|^2$, using **A**1 and Grönwall's inequality, we get $\|\xi_s^x(\theta)\|^2 \leq \mathrm{e}^{-2mt} \|x\|^2$ which implies the result for $\ell = 2$. Let now $\ell > 2$. Using again Lemma 21-a), Faà di Bruno's formula [30, Theorem 1] and since (19) can be written on the form

$$\frac{\mathrm{d}\varphi_s(\theta)}{\mathrm{d}s}\Big|_{s=t} = -\sum_{j=1}^{d} f'(\varphi_t(\theta)) \{\mathbf{e}_j\} \mathbf{e}_j ,$$

for all $i \in \{2, \ldots, \ell\}$, $\theta \in \mathbb{R}^d$ and $x_1, \cdots, x_i \in \mathbb{R}^d$, the function $\xi_t^{x_1,\cdots,x_i}(\theta) = \varphi_t^{(i)}(\theta)\{x_1 \otimes \cdots \otimes x_i\}$ satisfies the ordinary differential equation:

$$\frac{\mathrm{d}\xi_s^{x_1,\cdots,x_i}(\theta)}{\mathrm{d}s}\Big|_{s=t} = -\sum_{j=1}^{d} \sum_{\Omega \in \mathsf{P}(\{1,\ldots,i\})} f^{(|\Omega|+1)}(\varphi_t(\theta)) \left\{ \mathbf{e}_j \otimes \bigotimes_{l=1}^{i} \bigotimes_{j_1,\cdots,j_l \in \Omega} \xi_t^{x_{j_1},\cdots,x_{j_l}}(\theta) \right\} \mathbf{e}_j \;, \tag{55}$$

where $\mathsf{P}(\{1, \ldots, i\})$ is the set of partitions of $\{1, \ldots, i\}$, which does not contain the empty set and $|\Omega|$ is the cardinal of $\Omega \in \mathsf{P}(\{1, \ldots, i+1\})$. We now show by induction on i that for all $i \in \{1, \ldots, \ell\}$, there exists a universal constant C_i such that for all $t \geq 0$ and $\theta \in \mathbb{R}^d$,

$$\sup_{x \in \mathbb{R}^d} \left\| \varphi_t^{(i)}(\theta) \right\| \leq C_i \mathrm{e}^{-\mu t} \;. \tag{56}$$

For $i = 1$, the result follows from the case $\ell = 1$. Assume that the result is true for $\{1, \ldots, i\}$ for $i \in \{1, \ldots, \ell - 1\}$. We show the result for $i+1$. By (55), we have for all $\theta \in \mathbb{R}^d$ and $x_1, \cdots, x_i \in \mathbb{R}^d$,

$$\frac{\mathrm{d}\left\|\xi_s^{x_1,\cdots,x_{i+1}}(\theta)\right\|^2}{\mathrm{d}s}\Big|_{s=t} = -\sum_{\Omega \in \mathsf{P}(\{1,\ldots,i+1\})} f^{(|\Omega|+1)}(\varphi_t(\theta)) \left\{ \xi_t^{x_1,\cdots,x_{i+1}}(\theta) \otimes \bigotimes_{l=1}^{i+1} \bigotimes_{j_1,\ldots,j_l \in \Omega} \xi_t^{x_{j_1},\cdots,x_{j_l}}(\theta) \right\} .$$

Isolating the term corresponding to $\Omega = \{\{1, \ldots, i+1\}\}$ in the sum above and using Young's inequality, **A**1, Grönwall's inequality and the induction hypothesis, we get that there exists a universal constant C_{i+1} such that for all $t \geq 0$ and $x \in \mathbb{R}^d$ (56) holds for $i+1$.

b) The proof is a consequence of a), (51), **A**6(ℓ, p) and Lebesgue's dominated convergence theorem.

□

6.6.2. *Proof of Theorem 7.* We preface the proof of the Theorem by two fundamental first estimates.

THEOREM 24. *Let $g : \mathbb{R}^d \to \mathbb{R}$ satisfying **A**6$(3, p)$ for $p \in \mathbb{N}$. Assume **A**1-**A**2-**A**3-**A**5. Furthermore, suppose that there exists $q \in \mathbb{N}$ and $C \geq 0$ such that for all $\theta \in \mathbb{R}^d$,*

$$\mathbb{E}\left[\|\varepsilon_1(\theta)\|^{p+3}\right] \leq C(1 + \|\theta - \theta^*\|^q) \;,$$

and $\boldsymbol{A}$*4*$(2\tilde{p})$ *holds for* $\tilde{p} = p + 3 + q \vee k_\varepsilon$. *Let* $C_{\tilde{p}}$ *be the numerical constant given by Lemma 13 associated with* $\tilde{p}$.

(a) *For all* $\gamma \in (0, 1/(LC_{\tilde{p}}))$, $k \in \mathbb{N}^*$, *and starting point* $\theta_0 \in \mathbb{R}^d$,

$$\mathbb{E}\left[k^{-1}\sum_{i=1}^{k}\left\{g(\theta_i^{(\gamma)}) - g(\theta^*)\right\}\right] = \frac{h_g(\theta_0) - \mathbb{E}\left[h_g(\theta_{k+1}^{(\gamma)})\right]}{k\gamma}$$
$$+(\gamma/2)\int_{\mathbb{R}^d} h_g''(\tilde{\theta})\mathbb{E}\left[\left\{\varepsilon_1(\tilde{\theta})\right\}^{\otimes 2}\right]\mathrm{d}\pi_\gamma(\tilde{\theta}) - (\gamma/k)\tilde{A}_1(\theta_0, k) - \gamma^2\tilde{A}_2(\theta_0, k)\ ,$$

where $\theta_k^{(\gamma)}$ *is the Markov chain starting from* θ_0, *defined by the recursion* (1), *and*

$$\sup_{i \in \mathbb{N}^*}\tilde{A}_1(\theta_0, i) \le C\left\{1 + \|\theta_0 - \theta^*\|^{\tilde{p}}\right\}\ , \tag{57}$$

$$\tilde{A}_2(\theta_0, k) \le C\left\{1 + \|\theta_0 - \theta^*\|^{\tilde{p}}/k\right\}\ , \tag{58}$$

for some constant $C \ge 0$ *independent of* γ *and* k.

(b) *For all* $\gamma \in (0, 1/(LC_{\tilde{p}}))$,

$$\left|\int_{\mathbb{R}^d} g(\tilde{\theta})\pi_\gamma(\mathrm{d}\tilde{\theta}) - g(\theta^*) + (\gamma/2)\int_{\mathbb{R}^d} h_g''(\tilde{\theta})\mathbb{E}\left[\left\{\varepsilon(\tilde{\theta})\right\}^{\otimes 2}\right]\mathrm{d}\pi_\gamma(\tilde{\theta})\right| \le C\gamma^2\ .$$

PROOF. (a) Let $k \in \mathbb{N}^*$, $\gamma > 0$ and $\theta \in \mathbb{R}^d$. Consider the sequence $(\theta_k^{(\gamma)})_{k \ge 0}$ defined by the stochastic gradient recursion (1) and starting at θ. Theorem 23-b) shows that $h_g \in \mathrm{C}^3(\mathbb{R}^d)$. Therefore using (1) and the Taylor expansion formula, we have for all $i \in \{1, \ldots, k\}$

$$\begin{aligned} h_g(\theta_{i+1}^{(\gamma)}) = h_g(\theta_i^{(\gamma)}) &+ \gamma {h_g}'(\theta_i^{(\gamma)})\left\{-f'(\theta_i^{(\gamma)}) + \varepsilon_{i+1}(\theta_i^{(\gamma)})\right\} \\ &+ (\gamma^2/2)h_g''(\theta_i^{(\gamma)})\left\{-f'(\theta_i^{(\gamma)}) + \varepsilon_{i+1}(\theta_i^{(\gamma)})\right\}^{\otimes 2} \\ &+ (\gamma^3/(3!))h_g^{(3)}(\theta_i^{(\gamma)} + s_i^{(\gamma)}\Delta\theta_{i+1}^{(\gamma)})\left\{-f'(\theta_i^{(\gamma)}) + \varepsilon_{i+1}(\theta_i^{(\gamma)})\right\}^{\otimes 3}\ , \end{aligned}$$

where $s_i^{(\gamma)} \in [0,1]$ and $\Delta\theta_{i+1}^{(\gamma)} = \theta_{i+1}^{(\gamma)} - \theta_i^{(\gamma)}$. Therefore by (52), we get

$$
\begin{aligned}
k^{-1}\sum_{i=1}^{k}\left\{g(\theta_i^{(\gamma)}) - g(\theta^*)\right\} &= \frac{h_g(\theta) - h_g(\theta_{k+1}^{(\gamma)})}{k\gamma} + k^{-1}\sum_{i=1}^{k} {h_g}'(\theta_{i-1}^{(\gamma)})\varepsilon_{i+1}(\theta_i^{(\gamma)}) \\
&+(\gamma/(2k))\sum_{i=1}^{k} h_g''(\theta_i^{(\gamma)})\left\{-f'(\theta_i^{(\gamma)}) + \varepsilon_{i+1}(\theta_i^{(\gamma)})\right\}^{\otimes 2} \\
&+(\gamma^2/(3!k))\sum_{i=1}^{k} h_g^{(3)}(\theta_i^{(\gamma)} + s_i^{(\gamma)}\Delta\theta_{i+1}^{(\gamma)})\left\{-f'(\theta_i^{(\gamma)}) + \varepsilon_{i+1}(\theta_i^{(\gamma)})\right\}^{\otimes 3} .
\end{aligned}
$$

Taking the expectation and using **A**3, we have

$$
\begin{aligned}
\mathbb{E}\left[k^{-1}\sum_{i=1}^{k}\left\{g(\theta_i^{(\gamma)}) - g(\theta^*)\right\}\right] &= \frac{\mathbb{E}\left[h_g(\theta) - h_g(\theta_{k+1}^{(\gamma)})\right]}{k\gamma} \\
+(\gamma/2)\int_{\mathbb{R}^d} h_g''(\tilde{\theta})\mathbb{E}\left[\left\{\varepsilon_1(\tilde{\theta})\right\}^{\otimes 2}\right] \mathrm{d}\pi_\gamma(\tilde{\theta}) &- (\gamma/(2k))\tilde{B}_1 + (\gamma^2/(3!k))\tilde{B}_2 ,
\end{aligned}
$$

where

$$
\begin{aligned}
\tilde{B}_1(\theta_0,k) &= \mathbb{E}\left[\sum_{i=1}^{k}\left(h_g''(\theta^*)\left\{\varepsilon_1(\theta^*)\right\}^{\otimes 2} - h_g''(\theta_i^{(\gamma)})\left\{-f'(\theta_i^{(\gamma)}) + \varepsilon_{i+1}(\theta_i^{(\gamma)})\right\}^{\otimes 2}\right)\right] \\
\tilde{B}_2(\theta_0,k) &= \mathbb{E}\left[\sum_{i=1}^{k} h_g^{(3)}(\theta_i^{(\gamma)} + s_i^{(\gamma)}\Delta\theta_{i+1}^{(\gamma)})\left\{-f'(\theta_i^{(\gamma)}) + \varepsilon_{i+1}(\theta_i^{(\gamma)})\right\}^{\otimes 3}\right] .
\end{aligned}
$$

Then it remains to show that (57) and (58) holds. By **A**2, Theorem 7-b) and **A**5, there exists $C \geq 0$ such that we have that for all $\theta \in \mathbb{R}^d$,

$$
\left\|H'(\theta)\right\| \leq C_1(1 + \|\theta - \theta^*\|^{k_\varepsilon + p + 2}) ,
$$

where $H : \theta \mapsto h_g''(\theta)\mathbb{E}[\{-f'(\theta) + \varepsilon_1(\theta)\}^{\otimes 2}]$. Therefore (57) follows from **A**3, Lemma 15 and Proposition 16. Finally by Theorem 23-b) and Jensen inequality, there exists $C \geq 0$ such that for all $i \in \{1, \ldots, k\}$, almost surely,

$$
\begin{aligned}
& h_g^{(3)}(\theta_i^{(\gamma)} + s_i^{(\gamma)}\Delta\theta_{i+1}^{(\gamma)})\left\{-f'(\theta_i^{(\gamma)}) + \varepsilon_{i+1}(\theta_i^{(\gamma)})\right\}^{\otimes 3} \\
&\quad \leq C\left(1 + \left\|\theta_i^{(\gamma)}\right\|^{p_2} + \left\|\varepsilon_{i+1}(\theta_i^{(\gamma)})\right\|^{p_2}\right)\left(\left\|f'(\theta_i^{(\gamma)})\right\|^3 + \left\|\varepsilon_{i+1}(\theta_i^{(\gamma)})\right\|^3\right) .
\end{aligned}
$$

The proof of (58) then follows from **A**2, **A**3, (57) and Lemma 13.

(b) This result is a direct consequence of Proposition 16 and (a).

□

Proof of Theorem 7. Under the stated assumptions, the functions $\psi : \theta \mapsto h''_g(\theta)\mathbb{E}[\{\varepsilon(\theta)\}^{\otimes 2}]$ and g satisfy the conditions of Theorem 24. The proof then follows from combining Theorem 24-(b) applied to ψ and Theorem 24 applied to g. □

7. Acknowledgments. The authors would like to thank Éric Moulines and Arnak Dalalyan for helpful discussions. We acknowledge support from the European Research Council (grant SEQUOIA 724063). We acknowledge support from the chaire Economie des nouvelles donnees with the data science joint research initiative with the fonds AXA pour la recherche, and the Initiative de Recherche "Machine Learning for Large-Scale Insurance" from the Institut Louis Bachelier.

REFERENCES

[1] A. Abdulle, G. Vilmart, and K. C. Zygalakis. High order numerical approximation of the invariant measure of ergodic SDEs. *SIAM J. Numer. Anal.*, 52(4):1600–1622, 2014.

[2] R. Aguech, E. Moulines, and P. Priouret. On a perturbation approach for the analysis of stochastic tracking algorithms. *SIAM J. Control Optim.*, 39(3):872–899, 2000.

[3] F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *J. Mach. Learn. Res.*, 15(1):595–627, January 2014.

[4] F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pages 451–459, USA, 2011. Curran Associates Inc.

[5] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[6] M. Benaim. A dynamical system approach to stochastic approximations. *SIAM J. Control Optim.*, 34(2):437–472, 1996.

[7] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1990. Translated from the French by Stephen S. Wilson.

[8] D. Bertsekas. *Nonlinear programming.* Athena Scientific, 1995.

[9] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[10] C. Bouton and G. Pagès. About the multidimensional competitive learning vector quantization algorithm with constant gain. *Ann. Appl. Probab.*, 7(3):679–710, 1997.

[11] J. Chee and P. Toulis. Convergence diagnostics for stochastic gradient descent with constant step size. *arXiv preprint arXiv:1710.06382*, 2017.

[12] C. Chen, N. Ding, and L. Carin. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *NIPS*, pages 2269–2277, 2015.

[13] Xi Chen, Jason D Lee, Xin T Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *arXiv preprint arXiv:1610.08637*, 2016.

[14] A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

[15] A. Défossez and F. Bach. Averaged least-mean-squares: bias-variance trade-offs and optimal sampling distributions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics, (AISTATS)*, 2015.

[16] A. Dieuleveut and F. Bach. Nonparametric stochastic approximation with large step-sizes. *Ann. Statist.*, 44(4):1363–1399, 08 2016.

[17] A. Dieuleveut, A. Durmus, and Bach F. Supplement to bridging the gap between constant step size stochastic gradient descent and Markov chains, 2018.

[18] A. Dieuleveut, N. Flammarion, and F. Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *ArXiv e-prints*, February 2016.

[19] A. Durmus, U. Şimşekli, E. Moulines, R. Badeau, and G. Richard. Stochastic gradient Richardson-Romberg Markov Chain Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 2047–2055, 2016.

[20] A. Durmus and É. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 06 2017.

[21] J.-C. Fort and G. Pagès. Convergence of stochastic algorithms: from the Kushner-Clark theorem to the Lyapounov functional method. *Adv. in Appl. Probab.*, 28(4):1072–1094, 1996.

[22] J.-C. Fort and G. Pagès. Asymptotic behavior of a Markovian stochastic algorithm with constant step. *SIAM J. Control Optim.*, 37(5):1456–1482, 1999.

[23] M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems*. Number vol. 260 in Grundlehren der mathematischen Wissenschaften. Springer, 1998.

[24] P. Glynn and S. Meyn. A Liapunov bound for solutions of the Poisson equation. *Ann. Probab.*, 24(2):916–931, 04 1996.

[25] P. Hartman. *Ordinary Differential Equations: Second Edition*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 1982.

[26] P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Parallelizing stochastic approximation through mini-batching and tail-averaging. *ArXiv e-prints*, October 2016.

[27] P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Accelerating Stochastic Gradient Descent. *arXiv preprint arXiv:1704.08227*, 2017.

[28] H. J. Kushner and D. S. Clark. *Stochastic approximation methods for constrained and unconstrained systems*, volume 26 of *Applied Mathematical Sciences*. Springer-Verlag, New York-Berlin, 1978.

[29] G. Lan. An optimal method for stochastic composite optimization. *Math. Program.*, 133(1-2, Ser. A):365–397, 2012.

[30] E. Levy. Why do partitions occur in Faa di Bruno's chain rule for higher derivatives? Technical Report 0602183, arXiv, February 2006.

[31] L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Trans. Automatic Control*, AC-22(4):551–575, 1977.

[32] L. Ljung, G. Pflug, and H. Walk. *Stochastic approximation and optimization of random systems*. DMV Seminar. Birkhauser Verlag, Basel, Boston, 1992.

[33] S. Mandt, M. Hoffman, and D. M. Blei. A variational analysis of stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 354–363, 2016.

[34] S. Mandt, M. D Hoffman, and D. M. Blei. Stochastic gradient descent as approximate Bayesian inference. *arXiv preprint arXiv:1704.04289*, 2017.

[35] J. C. Mattingly, A. M. Stuart, and D. J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Process. Appl.*, 101(2):185–232, 2002.

[36] M. Métivier and P. Priouret. Applications of a Kushner and Clark lemma to general classes of stochastic algorithms. *IEEE Trans. Inform. Theory*, 30(2, part 1):140–151, 1984.

[37] M. Métivier and P. Priouret. Théorèmes de convergence presque sure pour une classe d'algorithmes stochastiques à pas décroissant. *Probab. Theory Related Fields*, 74(3):403–428, 1987.

[38] S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

[39] E. Moulines, P. Priouret, and F. Roueff. On recursive estimation for time varying autoregressive processes. *Ann. Statist.*, 33(6):2610–2654, 2005.

[40] A. Nedić and D. Bertsekas. Convergence rate of incremental subgradient algorithms. In *Stochastic optimization: algorithms and applications*, pages 223–264. Springer, 2001.

[41] D. Needell, R. Ward, and N. Srebro. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. In *Advances in Neural Information Processing Systems 27*, pages 1017–1025. Curran Associates, Inc., 2014.

[42] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM J. on Optimization*, 19(4):1574–1609, January 2009.

[43] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

[44] Y. Nesterov and J. Ph. Vial. Confidence level solutions for stochastic programming. *Automatica*, 44(6):1559–1568, June 2008.

[45] G. Pflug. Stochastic minimization with constant step-size: asymptotic laws. *SIAM Journal on Control and Optimization*, 24(4):655–666, 1986.

[46] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, 1992.

[47] P. Priouret and A. Veretenikov. A remark on the stability of the l.m.s. tracking algorithm. *Stochastic Analysis and Applications*, 16(1):119–129, 1998.

[48] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *ArXiv e-prints*, September 2011.

[49] H. Robbins and S. Monro. A stochastic approxiation method. *The Annals of mathematical Statistics*, 22(3):400–407, 1951.

[50] D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

[51] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2009.

[52] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 807–814, New York, NY, USA, 2007. ACM.

[53] O. Shamir and T. Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. *Proceedings of the $30^t h$ International Conference on Machine Learning*, 2013.

[54] J. Stoer and R. Bulirsch. *Introduction to numerical analysis*, volume 12. Springer Science & Business Media, 2013.

[55] Weijie J Su and Yuancheng Zhu. Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent. *arXiv preprint arXiv:1802.04876*, 2018.

[56] V. B. Tadić and A. Doucet. Asymptotic bias of stochastic gradient search. *Ann. Appl. Probab.*, 27(6):3255–3304, 2017.

[57] D. Talay and L. Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Anal. Appl.*, 8(4):483–509 (1991), 1990.

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

[59] M. Welling and Y. W Teh. Bayesian learning via Stochastic Gradient Langevin Dynamics. In *ICML*, pages 681–688, 2011.

[60] D. L. Zhu and P. Marcotte. Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities. *SIAM Journal on Optimization*, 6(3):714–726, 1996.

INRIA - Département d'informatique de l'ENS, École normale supérieure,
CNRS, PSL Research University 75005 Paris, France
E-mail: *aymeric.dieuleveut@inria.fr
E-mail: †francis.bach@inria.fr
CMLA - École normale supérieure Paris-Saclay,
94235 Cachan, France E-mail: **alain.durmus@cmla.ens-cachan.fr