The Annals of Statistics
2020, Vol. 48, No. 3, 1348–1382
https://doi.org/10.1214/19-AOS1850

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

BY AYMERIC DIEULEVEUT[1], ALAIN DURMUS[2] AND FRANCIS BACH[3]

[1]*CMAP, École Polytechnique (X), aymeric.dieuleveut@polytechnique.edu*
[2]*CMLA - École normale supérieure Paris-Saclay, CNRS, Université Paris-Saclay, alain.durmus@cmla.ens-cachan.fr*
[3]*INRIA - Département d'informatique de l'ENS, École normale supérieure, CNRS, PSL Research University, francis.bach@inria.fr*

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 (nonquadratic) 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, 53, 54], optimization [44, 46] and stochastic approximation [30, 48, 52]. The most widely used algorithms are stochastic gradient descent (SGD), a.k.a. Robbins–Monro algorithm [51], and some of its modifications based on averaging of the iterates [48, 50, 55].

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 $O(1/k)$ [51], but it appeared to be nonrobust to ill-conditioning and slower decays such as $O(1/\sqrt{k})$ together with averaging lead to both good practical and theoretical performance [3, 44].

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 the several choices of parameters to deal with decaying step sizes, for example, as $1/(\square k + \triangle)^\circ$; the initial conditions are forgotten exponentially fast[1] for well-conditioned (e.g., strongly convex) problems [42, 43], 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.

Received April 2018; revised April 2019.
MSC2010 subject classifications. Primary 62L20; secondary 90C15, 93E35.
Key words and phrases. Stochastic gradient descent, Markov chains.

[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 misspecification [54].

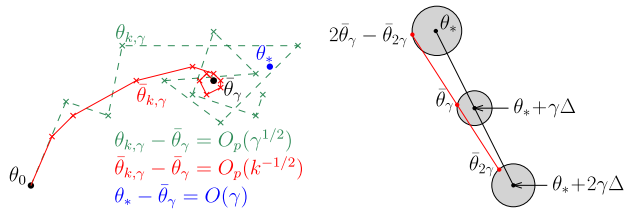

FIG. 1. *(Left) Convergence of iterates* $\theta_k^{(\gamma)}$ *and averaged iterates* $\bar{\theta}_k^{(\gamma)}$ *to the mean* $\bar{\theta}_\gamma$ *under the stationary distribution* π_γ. *(Right) Richardson–Romberg extrapolation, the disks are of radius* $O(\gamma^2)$.

We 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

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

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 nonaveraged iterates have an average magnitude of $\gamma^{1/2}$ [47].

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 $O(1/\sqrt{k})$ 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 [59], we achieve a nonasymptotic 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 [56]. 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.
- 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 nonasymptotic regimes. Moreover, convergence results can be used to derive confidence intervals for θ^*, as in [13, 57]. Another important application is the design of automatic restart schemes for SGD. In applications (especially in nonconvex settings), practitioners typically use epoch-wise constant step size; the step size is periodically reduced [26, 29]. 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 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 kth tensor power of E and $x^{\otimes k} \in E^{\otimes k}$ the kth 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 nth 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 $l \in \{1, \ldots, d\}$, the lth 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

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

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, that is, 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 $x \in \mathbb{R}^d$, δ_x stands for the Dirac measure at x.

2. Main results. In this section we describe the assumptions underlying our analysis, 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 $\mu > 0$, that is, for all $\theta_1, \theta_2 \in \mathbb{R}^d$ and $t \in [0, 1]$,

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

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

$$\big\|f'(\theta_1) - f'(\theta_2)\big\| \leq L\|\theta_1 - \theta_2\|.$$

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), $(\varepsilon_k)_{k\geq 1}$ is a sequence of random functions from $\mathbb{R}^d$ to $\mathbb{R}^d$ satisfying the following properties:

A3. There exists a filtration $(\mathcal{F}_k)_{k\geq 0}$ (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 $\mathbb{E}[\varepsilon_{k+1}(\theta)|\mathcal{F}_k] = 0$. In addition, $(\varepsilon_k)_{k\in\mathbb{N}^*}$ are 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', that is, 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 semistochastic. 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 $p \geq 2$:

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

Almost sure L-co-coercivity [62] is satisfied, for example, 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, $p \geq 2$, it suffices to show that f_1' is almost surely L-cocoercive and $\mathbb{E}^{1/p}[\|\varepsilon_1(\theta^*)\|^p] \leq \tau_p$. 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

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

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

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

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

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

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

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

then **A**4(4) is satisfied. This assumption is satisfied, for example, 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 selfconcordance 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 $k \geq 1$,

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

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

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

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 unaveraged 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

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

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 [49] which extends the classical Foster–Lyapunov criterion from Markov chain theory (see [40]) 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 nonquadratic 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 [33] and [30] and has been applied in numerous papers since then; see [6, 7, 38, 39, 58]. 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 [41] 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 [31]. Analysis has been extended to "tail averaging" [28] 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 [59] and [37] in the case of the Euler–Maruyama scheme. Then, [59] 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, [20] proposed and analyzed the use of Richardson–Romberg extrapolation applied to the stochastic gradient Langevin dynamics (SGLD) algorithm. This method introduced by [61] combines SGD and the Euler–Maruyama discretization of the Langevin diffusion associated to a target probability measure [14, 19]. 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 [35, 36], 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 [40] 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 $(\theta_k^{(\gamma)})_{k\in\mathbb{N}}$, that is, 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 $\mathcal{B}(\mathbb{R}^d)$ and 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 [60]. 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

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

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

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* $\theta \in \mathbb{R}^d$, $k \in \mathbb{N}^*$:

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

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

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

PROOF. Let $\gamma \in (0, 2/L)$ and $\lambda_1, \lambda_2 \in \mathcal{P}_2(\mathbb{R}^d)$. By [60], 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, that is, 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

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

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

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

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

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

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

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 [60], 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 $\pi_\gamma^{\lambda_1} \in \mathcal{P}_2(\mathbb{R}^d)$:

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

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

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

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 $k \in \mathbb{N}^*$,

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

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 $a, b \in \mathbb{R}^d$, $\|a - b\|^2 \le 2(\|a\|^2 + \|2\|^{[}]b)$ 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 $k \in \mathbb{N}^*$:

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

which concludes the proof of Proposition (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 Theorem 16 in Section 6.2.

3.1. *Expansion of moments of π_γ when γ is in a neighborhood of 0.* In this subsection 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 [34, 47] 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, that is, f' is affine, we have the following result:

PROPOSITION 3. *Assume $f = f_\Sigma$, $f_\Sigma : \theta \mapsto \|\Sigma^{1/2}(\theta - \theta^*)\|^2/2$, 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, for example, [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 [47] 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

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

$\bar{\theta}_\gamma$ *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 nonquadratic ($f'''(\theta^*) \neq 0$) 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 sub-section 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* $\theta_0 \in \mathbb{R}^d$, $k \in \mathbb{N}^*$,

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

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), (ii) a bias term which scales as $1/k^2$ and depends on the initial point $\theta_0 \in \mathbb{R}^d$ and (iii) a nonpositive 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

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

where we have used that

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

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, that is, with the notations of Example* 1, $f = f_\Sigma$, $\Sigma = \mathbb{E}[XX^\top]$, ε_k *are defined by* (6) *and step size* $\gamma \le 1/r^2$, *with* r *defined by* (8). *Assume* **A**1–**A**2–**A**3–**A**4(4). *For any starting*

point $\theta_0 \in \mathbb{R}^d$,

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

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

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

PROOF. The proof is postponed to the Supplementary Material [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 nonpositive 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), that is, $\theta_t = \theta^*$ 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

$$\begin{aligned}
\mathrm{D}(\mathcal{A}) &= \left\{h : \mathbb{R}^d \to \mathbb{R} : \text{ for all } \theta \in \mathbb{R}^d, \lim_{t\to 0} \frac{h(\varphi_t(\theta)) - h(\theta)}{t} \text{ exists } \right\}, \\
\mathcal{A}h(\theta) &= \lim_{t\to 0} \frac{\{h(\varphi_t(\theta)) - h(\theta)\}}{t} \quad \text{for all } h \in \mathrm{D}(\mathcal{A}), \theta \in \mathbb{R}^d.
\end{aligned} \tag{20}$$

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 coordinate-wise), 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 $i \in \{1, \dots, \ell\}$, $\|g^{(i)}(\theta)\| \le a_g\{\|\theta - \theta^*\|^p + b_g\}$.

THEOREM 7. *Let $g : \mathbb{R}^d \to \mathbb{R}$, satisfying* **A6**$(5, p)$ *for $p \in \mathbb{N}$. Assume* **A1**–**A2**–**A3**–**A5**. *Furthermore, suppose that there exists $q \in \mathbb{N}$ and $C \ge 0$ such that for all $\theta \in \mathbb{R}^d$,*

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

and **A4**$(2\tilde{p})$ *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*:

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

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 $C \ge 0$ independent of γ and k, we have*

$$A_1(\theta_0) \le C\{1 + \|\theta_0 - \theta^*\|^{\tilde{p}}\}, \qquad A_2(\theta_0, k) \le C\{1 + \|\theta_0 - \theta^*\|^{\tilde{p}}/k\}.$$

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 $C_1, C_2(\theta_0) \ge 0$ such that $\mathbb{E}[k^{-1}\sum_{i=1}^k \|\theta_i^{(\gamma)} - \theta^*\|^{2q}] = C_1\gamma + C_2(\theta_0)/k + O(\gamma^2)$.

3.4. *Discussion.* Classical proofs of convergence rely on another decomposition, originally proposed by [45] and used in recent papers analyzing the averaged iterate [4]. In order to highlight the main difference, we here sketch the arguments of these decompositions, 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 $k \in \mathbb{N}$,

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

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

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

Thus,

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

Averaging over the first k iterates yields

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

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

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

Thus, reorganizing terms,

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

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

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

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.

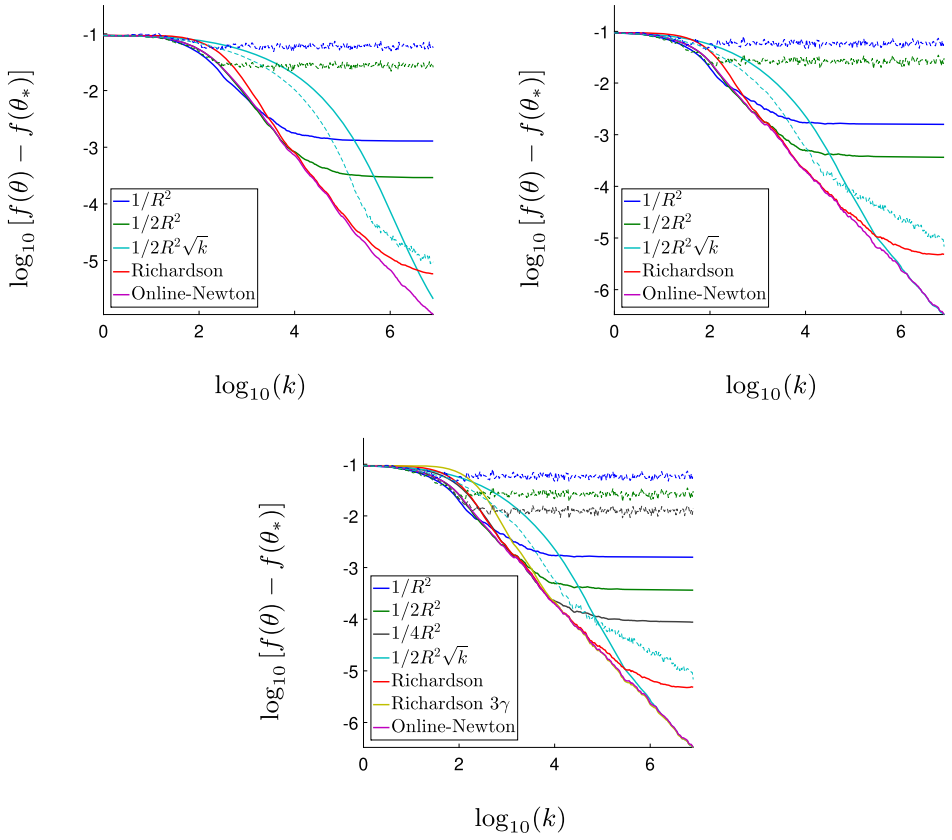

FIG. 2. *Synthetic data, logarithmic scales. Upper left*: *logistic regression*, $d = 12$, *with averaged SGD with step size* $1/R^2$, $1/2R^2$, *decaying step sizes* ($\gamma_k = 1/(2R^2\sqrt{k})$) (*averaged* (*plain*) *and nonaveraged* (*dashed*)), *Richardson–Romberg extrapolated iterates, and online 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 three different step sizes* 3γ, 2γ *and* $\gamma = 1/4R^2$.

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 unaveraged 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 three 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 (581,012 observations, $d = 54$), obtained from the LIBSVM data website.[3] Similarly, Richardson–Romberg iterates outper-

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

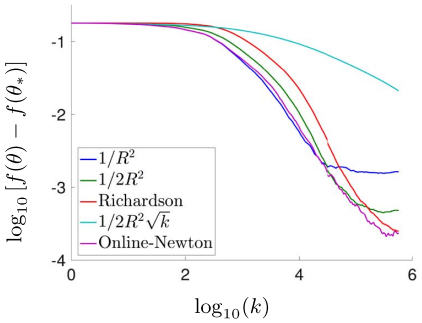

FIG. 3. *Covertype dataset.*

form constant step size while decaying step sizes are particularly slow. Convergence results are given in Figure 3.

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 selfconcordance condition [3].

6. Postponed proofs.

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 unbounded. 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 $q = 2, \ldots, p$,

$$\begin{aligned} &\mathbb{E}\big[\|f_1'(x) - f_1'(y)\|^q\big] \\ &\quad \leq L^{q-1}\|x - y\|^{q-2}\langle x - y, f'(x) - f'(y)\rangle, \end{aligned} \tag{25}$$

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

For Gaussian inputs Assumption **A**7 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 "semistochastic" 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 **A7**(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, so as to avoid the need for a constraint on the space where iterates live. It is straightforward to see that **A7**(p), $p \geq 2$, implies **A4**(p) with $\tau_p = \tilde{\tau}_p$, and **A8**–**A2** implies **A4**(4).

It is important to note that assuming **A3** (especially that $(\varepsilon_k)_{k \in \mathbb{N}^\star}$ are i.i.d. random fields) *does not* imply **A8**. On the contrary, making *the semi-stochastic assumption*, that is, 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 **A8**.

*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 **A7** 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* **A1**–**A2**–**A3**–**A4**(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 another 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):

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

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 another 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. □

LEMMA 9. *Assume* **A1**–**A2**–**A3**–**A4**(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 $\theta_0^{(\gamma)}$. □

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* $\gamma>0$,

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

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 [43], 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

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

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

$$\begin{aligned}&\mathbb{E}[\|f'_{k+1}(\theta_k^{(\gamma)})\|^2|\mathcal{F}_k]\\&\quad\le2(\mathbb{E}[\|f'_{k+1}(\theta_k^{(\gamma)})-f'_{k+1}(\theta^*)\|^2|\mathcal{F}_k]+\mathbb{E}[\|f'_{k+1}(\theta^*)\|^2|\mathcal{F}_k])\\&\quad\le2(\mathbb{E}[\|f'_{k+1}(\theta_k^{(\gamma)})-f'_{k+1}(\theta^*)\|^2|\mathcal{F}_k]+\tau^2)\\&\quad\le2(L\mathbb{E}[\langle f'_{k+1}(\theta_k^{(\gamma)})-f'_{k+1}(\theta^*),\theta_k^{(\gamma)}-\theta^*\rangle|\mathcal{F}_k]+\tau^2)\\&\quad\le2(L\langle f'(\theta_k^{(\gamma)})-f'(\theta^*),\theta_k^{(\gamma)}-\theta^*\rangle+\tau^2).\end{aligned}$$

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 $M\ge0$,

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

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 $M\ge0$,

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

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* $\theta \in \mathbb{R}^d$, $k \in \mathbb{N}^*$:

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

(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 Material [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* $\gamma \in (0, 2/(m + L)]$,

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

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

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

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* $\theta_0 \in \mathbb{R}^d$

$$\begin{aligned}&\mathbb{E}^{1/p}\big[\|\theta_k^{(\gamma)} - \theta^*\|^{2p}\big] \\ &\quad\leq \big(1 - 2\gamma\mu(1 - C_p\gamma L/2)\big)^k \mathbb{E}^{1/p}\big[\|2\|^{\lceil p\rceil}\big]\theta_0 - \theta^*] + \frac{D_p\gamma\tau_{2p}^2}{\mu},\end{aligned}$$

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*

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

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 Material [17], Section S3. □

LEMMA 15. *Let* $g : \mathbb{R}^d \to \mathbb{R}$ *satisfying* **A6**$(1, p)$ *for* $p \in \mathbb{N}$. *Then, for all* $\theta_1, \theta_2 \in \mathbb{R}^d$,

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

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

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

The proof is then concluded using **A6**(ℓ, p) and

$$\|\eta_s - \theta^*\| \leq \max\big(\|\theta_1 - \theta^*\|, \|\theta_2 - \theta^*\|\big). \qquad \square$$

PROPOSITION 16. *Let* $g : \mathbb{R}^d \to \mathbb{R}$, *satisfying* **A6**$(1, p)$ *for* $p \in \mathbb{N}$. *Assume* **A1**–**A2**–**A3**–**A4**$(2p)$. *Let* $C_p \geq 2$ *be given by Lemma 13 and only depending on* p. *For all* $\gamma \in (0, 1/(LC_p))$*and for all initial points* $\theta_0 \in \mathbb{R}^d$, *there exists* C_g *independent of* θ_0 *such that for all* $k \geq 1$,

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

PROOF. The proof is postponed to the Supplementary Material [17], Section S4. □

6.3. *Proof of Lemma 3.*

PROOF OF LEMMA 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 **A1** 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 $k = 1$,

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

Taking the expectation and using **A3**, $\theta_0^{(\gamma)}$ is independent of ε_1 and $\pi_\gamma R_\gamma = \pi_\gamma$, we get

$$\begin{aligned}
&\int_{\mathbb{R}^d} \big(\theta - \theta^*\big)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta) \\
&\quad = (\mathrm{Id} - \gamma\Sigma)\left[\int_{\mathbb{R}^d} \big(\theta - \theta^*\big)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta)\right](\mathrm{Id} - \gamma\Sigma) \\
&\qquad + \gamma^2 \int_{\mathbb{R}^d} \mathcal{C}(\theta)\pi_\gamma(\mathrm{d}\theta), \qquad (35) \\
&(\Sigma \otimes \mathrm{Id} + \mathrm{Id} \otimes \Sigma - \gamma\Sigma \otimes \Sigma)\left[\int_{\mathbb{R}^d} \big(\theta - \theta^*\big)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta)\right] \\
&\quad = \gamma \int_{\mathbb{R}^d} \mathcal{C}(\theta)\pi_\gamma(\mathrm{d}\theta).
\end{aligned}$$

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 $(\mathbf{f}_i)_{i\in\{0,\dots,d\}}$ as Σ. If we denote by $(\lambda_i)_{i\in\{0,\dots,d\}}$, 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 $\mathbb{R}^d \otimes \mathbb{R}^d$, $(\mathbf{f}_i \otimes \mathbf{f}_j)_{i,j\in\{0,\dots,d\}}$ and $(\lambda_i(1-\gamma\lambda_j) + \lambda_j(1-\gamma\lambda_i))_{i,j\in\{0,\dots,d\}}$ are its eigenvalues. □

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

PROPOSITION 17. *Assume that f is an objective function of a least-square regression problem, that is, 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 $\gamma \in (0, 1/r^2)$,*

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

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

PROOF. The proof follows the same line as the proof of Lemma 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,

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

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^*,

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

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, and we get

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

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 $(\mathbf{f}_i)_{i \in \{0,\dots,d\}}$ as $f''(\theta^*)$. If we denote by $(\lambda_i)_{i \in \{0,\dots,d\}}$, 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 $\mathbb{R}^d \otimes \mathbb{R}^d$, $(\mathbf{f}_i \otimes \mathbf{f}_j)_{i,j \in \{0,\dots,d\}}$ and $(\lambda_i(1 - \gamma\lambda_j) + \lambda_j(1 - \gamma\lambda_i))_{i,j \in \{0,\dots,d\}}$ 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 and 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 proven 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

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

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

$$
\begin{aligned}
&\int_{\mathbb{R}^d} (\theta-\theta^*)^{\otimes 3} \pi_\gamma(\mathrm{d}\theta) \\
&\quad = \gamma\mathbf{B}^{-1}\bigg[\int_{\mathbb{R}^d} \{\mathcal{C}(\theta)\} \otimes \{(\mathrm{Id}-\gamma f''(\theta^*))(\theta-\theta^*)\}\pi_\gamma(\mathrm{d}\theta)\bigg] \\
&\qquad + O(\gamma^2).
\end{aligned}
$$

By Lemma 19 we get

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

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

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

$$
\begin{aligned}
&\mathbb{E}\big[(\theta_1-\theta^*)^{\otimes 2}\big] \\
&\quad = \mathbb{E}\big[(\theta_0-\theta^*)^{\otimes 2} - \gamma(\mathrm{Id} \otimes f''(\theta^*) + f''(\theta^*) \otimes \mathrm{Id})(\theta-\theta^*)^{\otimes 2} \\
&\qquad + (\gamma/2)(\theta_0-\theta^*) \otimes \{f^{(3)}(\theta^*)(\theta_0-\theta^*)^{\otimes 2}\}
\end{aligned}
$$

$$+ (\gamma/2)\{f^{(3)}(\theta^*)(\theta_0-\theta^*)^{\otimes 2}\}\otimes(\theta_0-\theta^*) + \gamma^2\varepsilon_1(\theta_0)^{\otimes 2}(\theta_0)\big] + O(\gamma^3).$$

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

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

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 $\theta_0 \in \mathbb{R}^d$, $k \in \mathbb{N}^*$*

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

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

$$\begin{aligned}&\mathbb{E}\Big[\Big\{k^{-1}\sum_{i=0}^{k-1}\varphi(\theta_i^{(\gamma)})\Big\}^{\otimes 2}\Big]\\ &\quad= \frac{1}{k}\pi_\gamma[\psi_\gamma^{\otimes 2} - (\psi_\gamma-\varphi)^{\otimes 2}]\\ &\qquad- \frac{1}{k^2}[\pi_\gamma(\varpi_\gamma\varphi^\top + \varphi\varpi_\gamma^\top) + \chi_\gamma^2(\theta_0) - \chi_\gamma^1(\theta_0)] + O(k^{-3}),\end{aligned}$$

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; 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 $\Phi_k = k^{-1}\sum_{i=0}^{k-1}\varphi(\theta_i)$,

$$\begin{aligned}\mathbb{E}[\Phi_k] &= k^{-1}\sum_{i=0}^{k-1}\mathbb{E}[\varphi(\theta_i)]\\ &= k^{-1}\sum_{i=0}^{k-1}R_\gamma^i\varphi(\theta_0)\end{aligned}$$

$$= \pi_\gamma(\varphi) + k^{-1} \sum_{i=0}^{k-1} \big(R_\gamma^i \varphi(\theta_0) - \pi_\gamma(\varphi)\big))$$
$$= \pi_\gamma(\varphi) + k^{-1}\psi_\gamma(\theta_0) - R_\gamma^k \psi_\gamma(\theta_0) = \pi_\gamma(\varphi) + k^{-1}\psi_\gamma(\theta_0) + O\big(\rho^k\big).$$

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:

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

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

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

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

$$\|\varphi(R_\gamma^N\psi_\gamma)^\top(x) - \varphi(R_\gamma^N\psi_\gamma)^\top(y)\| \le C\rho^N\|x-y\|(1+\|x\|+\|y\|). \tag{45}$$

Then, we obtain by Lemma 11

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

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

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

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

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

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

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

$$
\begin{aligned}
&- \pi_\gamma\big[(\varphi-\psi_\gamma)(\varphi-\psi_\gamma)^\top - \psi_\gamma\psi_\gamma^\top\big]\big\} \\
&= \chi^2(\theta_0) - \chi^1(\theta_0).
\end{aligned}
$$

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:

A 9 (ℓ). $f \in C^\ell(\mathbb{R}^d)$ and there exists $M \geq 0$ such that for all $i \in \{2,\dots,\ell\}$, $\sup_{\theta\in\mathbb{R}^d} \|f^{(i)}(\theta)\| \leq \bar{L}$.

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*,

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

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* $i, j, l \in \{1,\dots,d\}$,

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

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

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

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

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

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

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

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

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

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

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

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($\ell$, p) *for ℓ, $p \in \mathbb{N}$, $\ell \geq 1$. Assume* **A**1 *and* **A**9($\ell + 1$).

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

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

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

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

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* **A**6$(\ell, p)$ *for $\ell, p \in \mathbb{N}$, $\ell \geq 2$. Assume* **A**1–**A**9$(\ell + 1)$.

(a) *For all $i \in \{1, \dots, \ell\}$, 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 $i \in \{0, \dots, \ell\}$, 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 $x \in \mathbb{R}^d$, $\theta \in \mathbb{R}^d$, $\xi_t^x(\theta) = D\varphi_t(\theta)\{x\}$ satisfies

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

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 [32], Theorem 1, and since (19) can be written on the form

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

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

$$\begin{aligned}\left.\frac{\mathrm{d}\xi_s^{x_1,\dots,x_i}(\theta)}{\mathrm{d}s}\right|_{s=t} = &-\sum_{j=1}^d \sum_{\Omega \in \mathsf{P}(\{1,\dots,i\})} f^{(|\Omega|+1)}\big(\varphi_t(\theta)\big) \\ &\times \left\{\mathbf{e}_j \otimes \bigotimes_{l=1}^{i} \bigotimes_{j_1,\dots,j_l \in \Omega} \xi_t^{x_{j_1},\dots,x_{j_l}}(\theta)\right\}\mathbf{e}_j,\end{aligned} \tag{55}$$

where $\mathsf{P}(\{1, \dots, i\})$ is the set of partitions of $\{1, \dots, i\}$, which does not contain the empty set, and $|\Omega|$ is the cardinal of $\Omega \in \mathsf{P}(\{1, \dots, i + 1\})$. We now show by induction on i that for all $i \in \{1, \dots, \ell\}$, 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 $\{1, \dots, i\}$ for $i \in \{1, \dots, \ell - 1\}$. We show the result for $i + 1$. By (55), we have for all $\theta \in \mathbb{R}^d$ and $x_1, \dots, x_i \in \mathbb{R}^d$,

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

Isolating the term corresponding to $\Omega = \{\{1, \dots, i + 1\}\}$ 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 $\theta \in \mathbb{R}^d$,*

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

and **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 $\theta_0 \in \mathbb{R}^d$,*

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

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

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

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

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

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

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

PROOF.

(a) Let $k \in \mathbb{N}^*$, $\gamma > 0$ and $\theta \in \mathbb{R}^d$. Consider the sequence $(\theta_k^{(\gamma)})_{k\geq 0}$, 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 $i \in \{1, \dots, k\}$

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

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

$$k^{-1}\sum_{i=1}^{k}\{g(\theta_i^{(\gamma)}) - g(\theta^*)\} = \frac{h_g(\theta) - h_g(\theta_{k+1}^{(\gamma)})}{k\gamma} + k^{-1}\sum_{i=1}^{k} h_g{}'(\theta_{i-1}^{(\gamma)})\varepsilon_{i+1}(\theta_i^{(\gamma)})$$

$$+ (\gamma/(2k)) \sum_{i=1}^{k} h_g''(\theta_i^{(\gamma)}) \{-f'(\theta_i^{(\gamma)}) + \varepsilon_{i+1}(\theta_i^{(\gamma)})\}^{\otimes 2}$$

$$+ (\gamma^2/(3!k)) \sum_{i=1}^{k} h_g^{(3)}(\theta_i^{(\gamma)} + s_i^{(\gamma)} \Delta\theta_{i+1}^{(\gamma)}) \{-f'(\theta_i^{(\gamma)}) + \varepsilon_{i+1}(\theta_i^{(\gamma)})\}^{\otimes 3}.$$

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

$$\mathbb{E}\left[k^{-1} \sum_{i=1}^{k} \{g(\theta_i^{(\gamma)}) - g(\theta^*)\}\right]$$

$$= \frac{\mathbb{E}[h_g(\theta) - h_g(\theta_{k+1}^{(\gamma)})]}{k\gamma}$$

$$+ (\gamma/2) \int_{\mathbb{R}^d} h_g''(\tilde{\theta}) \mathbb{E}[\{\varepsilon_1(\tilde{\theta})\}^{\otimes 2}] \, \mathrm{d}\pi_\gamma(\tilde{\theta}) - (\gamma/(2k)) \tilde{B}_1 + (\gamma^2/(3!k)) \tilde{B}_2,$$

where

$$\tilde{B}_1(\theta_0, k) = \mathbb{E}\left[\sum_{i=1}^{k} (h_g''(\theta^*) \{\varepsilon_1(\theta^*)\}^{\otimes 2} - h_g''(\theta_i^{(\gamma)}) \{-f'(\theta_i^{(\gamma)}) + \varepsilon_{i+1}(\theta_i^{(\gamma)})\}^{\otimes 2})\right],$$

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

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 $\theta \in \mathbb{R}^d$,

$$\|H'(\theta)\| \leq C_1(1 + \|k\|_\varepsilon^{[} + p + 2]\theta - \theta^*),$$

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 Theorem 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,

$$h_g^{(3)}(\theta_i^{(\gamma)} + s_i^{(\gamma)} \Delta\theta_{i+1}^{(\gamma)}) \{-f'(\theta_i^{(\gamma)}) + \varepsilon_{i+1}(\theta_i^{(\gamma)})\}^{\otimes 3}$$

$$\leq C(1 + \|\theta_i^{(\gamma)}\|^{p_2} + \|\varepsilon_{i+1}(\theta_i^{(\gamma)})\|^{p_2})(\|f'(\theta_i^{(\gamma)})\|^3 + \|3\|^{[}\varepsilon_{i+1}(\theta_i^{(\gamma)})).$$

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

(b) This result is a direct consequence of Theorem 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. □

Acknowledgements. The authors would like to thank Éric Moulines and Arnak Dalalyan for helpful discussions.

This work was supported by the European Research Council (grant SEQUOIA 724063).

This work was supported by 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.

SUPPLEMENTARY MATERIAL

Supplement to "Bridging the gap between constant step size stochastic gradient descent and Markov chains" (DOI: 10.1214/19-AOS1850SUPP; .pdf). Proofs of Lemma 11, Lemma 12, Lemma 13, Proposition 16 and Corollary 6.

REFERENCES

[1] ABDULLE, A., VILMART, G. and ZYGALAKIS, K. C. (2014). High order numerical approximation of the invariant measure of ergodic SDEs. *SIAM J. Numer. Anal.* **52** 1600–1622. MR3229658 https://doi.org/10.1137/130935616

[2] AGUECH, R., MOULINES, E. and PRIOURET, P. (2000). On a perturbation approach for the analysis of stochastic tracking algorithms. *SIAM J. Control Optim.* **39** 872–899. MR1786334 https://doi.org/10.1137/S0363012998333852

[3] BACH, F. (2014). Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *J. Mach. Learn. Res.* **15** 595–627. MR3190851

[4] BACH, F. and MOULINES, E. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances on Neural Information Processing Systems* (*NIPS*) 451–459.

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

[6] BENAIM, M. (1996). A dynamical system approach to stochastic approximations. *SIAM J. Control Optim.* **34** 437–472. MR1377706 https://doi.org/10.1137/S0363012993253534

[7] BENVENISTE, A., MÉTIVIER, M. and PRIOURET, P. (1990). *Adaptive Algorithms and Stochastic Approximations*. *Applications of Mathematics* (*New York*) **22**. Springer, Berlin. Translated from the French by Stephen S. Wilson. MR1082341 https://doi.org/10.1007/978-3-642-75894-2

[8] BERTSEKAS, D. P. (1999). *Nonlinear Programming*, 2nd ed. *Athena Scientific Optimization and Computation Series*. Athena Scientific, Belmont, MA. MR3444832

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

[10] BOUTON, C. and PAGÈS, G. (1997). About the multidimensional competitive learning vector quantization algorithm with constant gain. *Ann. Appl. Probab.* **7** 679–710. MR1459266 https://doi.org/10.1214/aoap/1034801249

[11] CHEE, J. and TOULIS, P. (2017). Convergence diagnostics for stochastic gradient descent with constant step size. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, (*AISTATS*).

[12] CHEN, C., DING, N. and CARIN, L. (2015). On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *Advances on Neural Information Processing Systems* (*NIPS*) 2269–2277.

[13] CHEN, X., LEE, J. D., TONG, X. T. and ZHANG, Y. (2016). Statistical inference for model parameters in stochastic gradient descent. ArXiv preprint. Available at arXiv:1610.08637.

[14] DALALYAN, A. S. (2017). Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 651–676. MR3641401 https://doi.org/10.1111/rssb.12183

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

[16] DIEULEVEUT, A. and BACH, F. (2016). Nonparametric stochastic approximation with large step-sizes. *Ann. Statist.* **44** 1363–1399. MR3519927 https://doi.org/10.1214/15-AOS1391

[17] DIEULEVEUT, A., DURMUS, A. and BACH, F. (2019). Supplement to "Bridging the gap between constant step size stochastic gradient descent and Markov Chains." https://doi.org/10.1214/19-AOS1850SUPP.

[18] DIEULEVEUT, A., FLAMMARION, N. and BACH, F. (2017). Harder, better, faster, stronger convergence rates for least-squares regression. *J. Mach. Learn. Res.* **18** 101. MR3725440

[19] DURMUS, A. and MOULINES, É. (2017). Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.* **27** 1551–1587. MR3678479 https://doi.org/10.1214/16-AAP1238

[20] DURMUS, A., ŞIMŞEKLI, U., MOULINES, E., BADEAU, R. and RICHARD, G. (2016). Stochastic gradient Richardson–Romberg Markov chain Monte Carlo. In *Advances in Neural Information Processing Systems* 2047–2055.

[21] FORT, J.-C. and PAGÈS, G. (1996). Convergence of stochastic algorithms: From the Kushner–Clark theorem to the Lyapounov functional method. *Adv. in Appl. Probab.* **28** 1072–1094. MR1418247 https://doi.org/10.2307/1428165

[22] FORT, J.-C. and PAGÈS, G. (1999). Asymptotic behavior of a Markovian stochastic algorithm with constant step. *SIAM J. Control Optim.* **37** 1456–1482. MR1710228 https://doi.org/10.1137/S0363012997328610

[23] FREIDLIN, M. I. and WENTZELL, A. D. (1998). *Random Perturbations of Dynamical Systems*, 2nd ed. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **260**. Springer, New York. MR1652127 https://doi.org/10.1007/978-1-4612-0611-8

[24] GLYNN, P. W. and MEYN, S. P. (1996). A Liapounov bound for solutions of the Poisson equation. *Ann. Probab.* **24** 916–931. MR1404536 https://doi.org/10.1214/aop/1039639370

[25] HARTMAN, P. (2002). *Ordinary Differential Equations*: *Second Edition*. *Classics in Applied Mathematics* **38**. SIAM, Philadelphia, PA. Corrected reprint of the second (1982) edition [Birkhäuser, Boston, MA; MR0658490 (83e:34002)], With a foreword by Peter Bates. MR1929104 https://doi.org/10.1137/1.9780898719222

[26] HE, K., ZHANG, X., REN, S. and SUN, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778.

[27] JAIN, P., KAKADE, S. M., KIDAMBI, R., NETRAPALLI, P. and SIDFORD, A. (2018). Accelerating stochastic gradient descent. In *Proceedings of the International Conference on Learning Theory* (*COLT*).

[28] JAIN, P., KAKADE, S. M., KIDAMBI, R., NETRAPALLI, P. and SIDFORD, A. (2017). Parallelizing stochastic gradient descent for least squares regression: Mini-batching, averaging, and model misspecification. *J. Mach. Learn. Res.* **18** 223. MR3827111

[29] KRIZHEVSKY, A., SUTSKEVER, I. and HINTON, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (*NIPS*) 1097–1105.

[30] KUSHNER, H. J. and CLARK, D. S. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. *Applied Mathematical Sciences* **26**. Springer, New York-Berlin. MR0499560

[31] LAN, G. (2012). An optimal method for stochastic composite optimization. *Math. Program.* **133** 365–397. MR2921104 https://doi.org/10.1007/s10107-010-0434-y

[32] LEVY, E. (2006). Why do partitions occur in Faa di Bruno's chain rule for higher derivatives? Technical Report 0602183.

[33] LJUNG, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Trans. Automat. Control* **AC-22** 551–575. MR0465458 https://doi.org/10.1109/tac.1977.1101561

[34] LJUNG, L., PFLUG, G. and WALK, H. (1992). *Stochastic Approximation and Optimization of Random Systems*. *DMV Seminar* **17**. Birkhäuser, Basel. MR1162311 https://doi.org/10.1007/978-3-0348-8609-3

[35] MANDT, S., HOFFMAN, M. and BLEI, D. M. (2016). A variational analysis of stochastic gradient algorithms. In *International Conference on Machine Learning* 354–363.

[36] MANDT, S., HOFFMAN, M. D. and BLEI, D. M. (2017). Stochastic gradient descent as approximate Bayesian inference. *J. Mach. Learn. Res.* **18** 134. MR3763768

[37] MATTINGLY, J. C., STUART, A. M. and HIGHAM, D. J. (2002). Ergodicity for SDEs and approximations: Locally Lipschitz vector fields and degenerate noise. *Stochastic Process. Appl.* **101** 185–232. MR1931266 https://doi.org/10.1016/S0304-4149(02)00150-3

[38] MÉTIVIER, M. and PRIOURET, P. (1984). Applications of a Kushner and Clark lemma to general classes of stochastic algorithms. *IEEE Trans. Inform. Theory* **30** 140–151. MR0807052 https://doi.org/10.1109/TIT.1984.1056894

[39] MÉTIVIER, M. and PRIOURET, P. (1987). Théorèmes de convergence presque sure pour une classe d'algorithmes stochastiques à pas décroissant. *Probab. Theory Related Fields* **74** 403–428. MR0873887 https://doi.org/10.1007/BF00699098

[40] MEYN, S. and TWEEDIE, R. L. (2009). *Markov Chains and Stochastic Stability*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2509253 https://doi.org/10.1017/CBO9780511626630

[41] MOULINES, E., PRIOURET, P. and ROUEFF, F. (2005). On recursive estimation for time varying autoregressive processes. *Ann. Statist.* **33** 2610–2654. MR2253097 https://doi.org/10.1214/009053605000000624

[42] NEDIĆ, A. and BERTSEKAS, D. (2001). Convergence rate of incremental subgradient algorithms. In *Stochastic Optimization*: *Algorithms and Applications* (*Gainesville*, *FL*, 2000). *Appl. Optim.* **54** 223–264. Kluwer Academic, Dordrecht. MR1835501 https://doi.org/10.1007/978-1-4757-6594-6_11

[43] NEEDELL, D., WARD, R. and SREBRO, N. (2014). Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. In *Advances in Neural Information Processing Systems* (*NIPS*), 1017–1025.

[44] NEMIROVSKI, A., JUDITSKY, A., LAN, G. and SHAPIRO, A. (2008). Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* **19** 1574–1609. MR2486041 https://doi.org/10.1137/070704277

[45] NEMIROVSKY, A. S. and YUDIN, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. *A Wiley-Interscience Publication*. Wiley, New York. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics. MR0702836

[46] NESTEROV, Y. and VIAL, J.-P. (2008). Confidence level solutions for stochastic programming. *Automatica J. IFAC* **44** 1559–1568. MR2531843 https://doi.org/10.1016/j.automatica.2008.01.017

[47] PFLUG, G. C. (1986). Stochastic minimization with constant step-size: Asymptotic laws. *SIAM J. Control Optim.* **24** 655–666. MR0846373 https://doi.org/10.1137/0324039

[48] POLYAK, B. T. and JUDITSKY, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* **30** 838–855. MR1167814 https://doi.org/10.1137/0330046

[49] PRIOURET, P. and VERETENNIKOV, A. Y. (1998). A remark on the stability of the LMS tracking algorithm. *Stoch. Anal. Appl.* **16** 119–129. MR1603888 https://doi.org/10.1080/07362999808809521

[50] RAKHLIN, A., SHAMIR, O. and SRIDHARAN, K. (2011). Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the* 29*th International Conference on Machine Learning*.

[51] ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668 https://doi.org/10.1214/aoms/1177729586

[52] RUPPERT, D. (1988). Efficient estimations from a slowly convergent Robbins–Monro process Technical report, Cornell University Operations Research and Industrial Engineering.

[53] SHALEV-SHWARTZ, S., SHAMIR, O., SREBRO, N. and SRIDHARAN, K. (2009). Stochastic convex optimization. In *Proceedings of the International Conference on Learning Theory* (*COLT*).

[54] SHALEV-SHWARTZ, S., SINGER, Y. and SREBRO, N. (2007). Pegasos: Primal estimated sub-GrAdient SOlver for SVM. In *Proceedings of the International Conference on Machine Learning*, *ICML* 807–814.

[55] SHAMIR, O. and ZHANG, T. (2013). Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the International Conference on Machine Learning*.

[56] STOER, J. and BULIRSCH, R. (2002). *Introduction to Numerical Analysis*, 3rd ed. *Texts in Applied Mathematics* **12**. Springer, New York. MR1923481 https://doi.org/10.1007/978-0-387-21738-3

[57] SU, W. J. and ZHU, Y. (2018). Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent. ArXiv preprint. Available at arXiv:1802.04876.

[58] TADIĆ, V. B. and DOUCET, A. (2017). Asymptotic bias of stochastic gradient search. *Ann. Appl. Probab.* **27** 3255–3304. MR3737925 https://doi.org/10.1214/16-AAP1272

[59] TALAY, D. and TUBARO, L. (1990). Expansion of the global error for numerical schemes solving stochastic differential equations. *Stoch. Anal. Appl.* **8** 483–509. MR1091544 https://doi.org/10.1080/07362999008809220

[60] VILLANI, C. (2009). *Optimal Transport*: *Old and New*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **338**. Springer, Berlin. MR2459454 https://doi.org/10.1007/978-3-540-71050-9

[61] WELLING, M. and TEH, Y. W. (2011). Bayesian learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the International Conference on Machine Learning* (*ICML*) 681–688.

[62] ZHU, D. L. and MARCOTTE, P. (1996). Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities. *SIAM J. Optim.* **6** 714–726. MR1402202 https://doi.org/10.1137/S1052623494250415