Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains

Aymeric Dieuleveut, Alain Durmus, Francis Bach

Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains

Aymeric Dieuleveut[1], Alain Durmus[2], and Francis Bach[1]

[1]INRIA - Département d'informatique de l'ENS, École normale supérieure, CNRS, PSL Research University, 75005 Paris, France
[2]LTCI, Telecom ParisTech, Université Paris-Saclay, 75013, Paris, France.

July 19, 2017

Abstract

We consider the minimization of an objective function given access to unbiased estimates of its gradient through stochastic gradient descent (SGD) with constant step-size. While the detailed analysis was only performed for quadratic functions, we provide an explicit asymptotic expansion of the moments of the averaged SGD iterates that outlines the dependence on initial conditions, the effect of noise and the step-size, as well as the lack of convergence in the general (non-quadratic) case. For this analysis, we bring tools from Markov chain theory into the analysis of stochastic gradient and create new ones (similar but different from stochastic MCMC methods). 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 (Bottou and Bousquet, 2008; Shalev-Shwartz et al., 2009, 2007), optimization (Nemirovski et al., 2009; Nesterov and Vial, 2008), and stochastic approximation (Kushner and Yin, 2003; Polyak and Juditsky, 1992; Ruppert, 1988). The most widely used algorithms are stochastic gradient descent (SGD), a.k.a. Robbins-Monro algorithm (Robbins and Monro, 1951), and some of its modifications based on averaging of the iterates (Polyak and Juditsky, 1992; Rakhlin et al., 2011; Shamir and Zhang, 2013).

While the choice of the step-size may be done robustly in the deterministic case (see, e.g., Bertsekas, 1995), this remains a traditional theoretical and practical issue in the stochastic case. Indeed, early work suggested to use step-size decaying with the number k of iterations as $O(1/k)$ (Robbins and Monro, 1951), but it appeared to be non-robust to ill-conditioning and slower decays such as $O(1/\sqrt{k})$ together with averaging lead to both good practical and theoretical performance (Bach, 2014).

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 single parameter value to set as opposed to the several choices of parameters to deal with decaying step-sizes, e.g., as $1/(\square k + \triangle)^\circ$; the initial conditions are forgotten exponentially fast for well-conditioned (e.g., strongly convex) problems (Nedić and Bertsekas, 2001; Needell et al., 2014), and the performance, although not optimal, is sufficient in practice (in a machine learning set-up, being only 0.1% away from the optimal prediction often does not matter).

The main goals of this paper are (a) to gain a complete understanding of the properties of constant-step-size SGD in the strongly convex case, and (b) to propose provable improvements to get closer to the optimum when precision matters or in high-dimensional settings. We consider the

iterates of the SGD recursion on $\mathbb{R}^d$ defined starting from $\theta_0 \in \mathbb{R}^d$, for $k \geq 0$, and a step-size $\gamma > 0$ by

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

where f is the objective function to minimize (in machine learning the generalization performance), $\varepsilon_{k+1}(\theta_k^{(\gamma)})$ the zero-mean statistically independent noise (in machine learning, obtained from a single i.i.d. observation of a data point). Following Bach and Moulines (2013), we leverage the property that the sequence of iterates $(\theta_k^{(\gamma)})_{k\geq 0}$ is an *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 its unique stationary distribution π_γ (see Section 3.1) highlighting the facts that (a) initial conditions of the algorithms are forgotten quickly and (b) the algorithm does not converge to a point but oscillates around the mean of π_γ. See an illustration in Figure 1 (left). It is known that the oscillations of the non-averaged iterates have an average magnitude of $\gamma^{1/2}$ (Pflug, 1986).

Consider the average 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 θ_* the global optimum 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 (Bach and Moulines, 2013), 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 that highlights all dependencies on initial conditions and noise variance, as achieved for least-squares by Défossez and Bach (2015), with an explicit decomposition into "bias" and "variance" terms: the bias term characterizes how fast initial conditions are forgotten and thus is increasing in a well-chosen norm of $\theta_0 - \theta_*$; 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, we achieve a non-asymptotic weak error expansion in the step-size between π_γ and the Dirac at θ_*. Namely, we prove that for all functions $g : \mathbb{R}^d \to \mathbb{R}$, regular enough, $\int_{\mathbb{R}^d} g(\theta)\mathrm{d}\pi_\gamma(\theta) = g(\theta_*) + \gamma C + O(\gamma^2)$ for some $C \in \mathbb{R}$ independent of γ. Given this expansion, we can now use a very simple trick from numerical analysis, namely Richardson-Romberg extrapolation (Stoer and Bulirsch, 2013): 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 both averaged iterates $(\bar{\theta}_k^{(\gamma)})_{k\geq 0}$ and $(\bar{\theta}_k^{(2\gamma)})_{k\geq 0}$ will converge to $\bar{\theta}_\gamma$ and $\bar{\theta}_{2\gamma}$ respectively. Since $\bar{\theta}_\gamma = \theta_* + \Delta\gamma + O(\gamma^2)$ and $\bar{\theta}_{2\gamma} = \theta_* + 2\Delta\gamma + O(\gamma^2)$, for $\Delta \in \mathbb{R}^d$ independent of γ, the combined iterate $2\bar{\theta}_k^{(\gamma)} - \bar{\theta}_k^{(2\gamma)}$ will converge to a point which is $\theta_* + O(\gamma^2)$ and we have thus gained one order in the convergence rate. See illustration in Figure 1(right).

In summary, we make the following contributions:

- We provide in Section 2 an asymptotic expansions of the mean of the averaged SGD iterate that outlines 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 bring and adapt in Section 3 tools from analysis of discretization of diffusion processes into the one of SGD and create new ones. We believe that this analogy and the associated ideas have their own interest.
- We show in Section 4 empirical improvements of the extrapolation schemes.

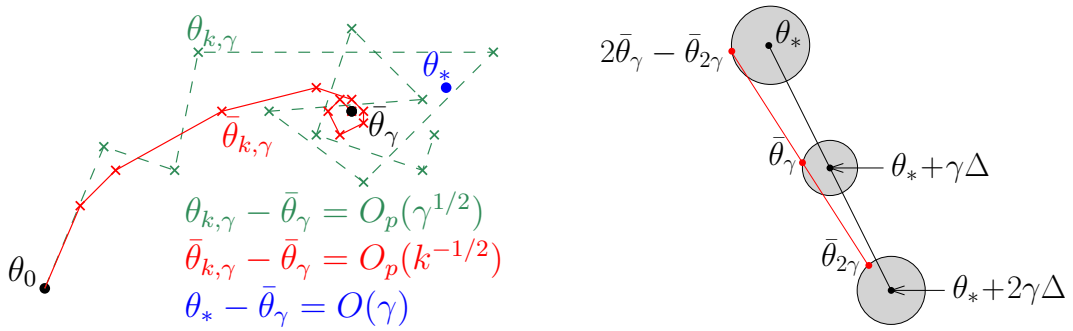

Figure 1: (Left) Convergence of iterates $\theta_k^{(\gamma)}$ and averaged iterates $\bar{\theta}_k^{(\gamma)}$ to the mean $\bar{\theta}^{(\gamma)}$ under the stationary distribution π_γ. (Right) Richardson-Romberg extrapolation, the disks are of radius $O(\gamma^2)$.

2 Main results

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

2.1 Setting

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

A 1. *The function f is strongly convex with convexity constant μ, i.e. $f - \frac{\mu}{2}\|\cdot\|^2$ is convex.*

A 2. *The function f is four times continuously differentiable with uniformly second to fourth bounded derivatives. Especially f is L-smooth: $\forall \theta \in \mathbb{R}^d$, the largest eigenvalue of $f''(\theta)$ is less than L.*

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

A 3. *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 $k \in \mathbb{N}$, for any $\theta \in \mathbb{R}^d$, $\varepsilon_{k+1}(\theta)$ is an $\mathcal{F}_{k+1}$-measurable random variable and $\mathbb{E}\left[\varepsilon_{k+1}(\theta)|\mathcal{F}_k\right] = 0$. In addition, $(\varepsilon_k)_{k\in\mathbb{N}^*}$ are independent and identically distributed (i.i.d.) random variables. Moreover, we assume that θ_0 is $\mathcal{F}_0$ measurable.*

A3 expresses that we observe a noisy gradient $f'_{k+1}(\theta_k^{(\gamma)}) = f'(\theta_k^{(\gamma)}) - \varepsilon_{k+1}(\theta_k^{(\gamma)})$ which are unbiased estimator of f'. Note that the notation f'_k does necessary presuppose the existence of functions f_k such that $(f_k)' = f'_k$. Note also that we do not assume that the random vectors $(\varepsilon_{k+1}(\theta_k^{(\gamma)}))_{k\in\mathbb{N}}$ are i.i.d., a stronger assumption generally referred to as the semi-stochastic setting. 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 $p \geq 2$:

A 4 (p). *ε_1 is almost surely L-co-coercive (with the same constant as in **A**2): for any $\eta, \theta \in \mathbb{R}^d$: $L\langle \varepsilon_1(\theta) - \varepsilon_1(\eta), \theta - \eta\rangle \geq \|\varepsilon_1(\theta) - \varepsilon_1(\eta)\|^2$. Moreover, there exists $\tau_p \geq 0$, such that $\varepsilon_1(\theta_*)$ admits bounded moments up to the order p: $\mathbb{E}^{1/p}[\|\varepsilon_1(\theta_*)\|^p] \leq \tau_p$.*

Almost sure L-co-coercivity (Zhu and Marcotte, 1996) is for example satisfied if there exist random functions f_k (such that $f'_k = (f_k)'$) which are a.s. convex and L-smooth. Note that a.s. co-coercive of the noise function ε_1 implies under **A**1, **A**2 the a.s. co-coercivity of the function f'_1. Weaker assumptions could be made on the noise (see Appendix A.3 for a discussion).

Learning from i.i.d. observations. Our main motivation comes from machine learning; namely, we consider sets $\mathcal{X}, \mathcal{Y}$, a convex loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}$. The objective function is the generalization error $f_\ell(\theta) = \mathbb{E}_{X,Y}[\ell(X, Y, \theta)]$. For any $k \geq 1$, we define $\varepsilon_k(\theta) = \ell(x_k, y_k, \theta) - f_\ell(\theta)$ which corresponds to following the negative gradient of a single i.i.d. observation $(x_k, y_k)_{k\geq 1}$; Assumption **A**3 is then satisfied with $\mathcal{F}_k := \sigma((x_j, y_j)_{1\leq j\leq k})$.

Two classical situations are worth mentioning: in *least-squares regression*, $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, and the loss function is $\ell(X, Y, \theta) = (\langle X, \theta \rangle - Y)^2$. Then f_ℓ is a quadratic function f_Σ, with $\Sigma = \mathbb{E}[XX^\top]$, thus satisfies Assumption **A**2. For any $p \geq 2$, Assumption **A**4(p) is satisfied as soon as the iterates are a.s. bounded, while **A**1 is satisfied if the second moment matrix is invertible or additional regularization is added. In this setting, ε_k can be decomposed as $\varepsilon_k = \varrho_k + \xi_k$ where ϱ_k is the multiplicative part, ξ_k the additive part, given for $\theta \in \mathbb{R}^d$ by $\varrho_k(\theta) = (x_k x_k^\top - \Sigma)(\theta - \theta_*)$ and

$$\xi_k = (x_k^\top \theta_* - y_k) x_k \ . \tag{4}$$

Note that for all $k \geq 1$, ξ_k does not depend on θ. This two parts in the noise will appear in Corollary 4. In *logistic regression*, where $\ell(X, Y, \theta) = \log(1 + \exp(-Y \langle X, \theta \rangle))$. Assumptions **A**4 or **A**2 are similarly satisfied, while **A**1 needs an additional restriction to a compact set. Using self-concordance assumptions (Bach, 2014) would allow a direct unconstrained application.

2.2 Related work

Constant step-size SGD. Several attempts have been made to improve convergence of SGD. Bach and Moulines (2013) propose an online Newton algorithm which converges to the optimal point with constant steps. While it behaves very well in practice, this algorithm has no convergence guarantees.

The quadratic case was studied by Bach and Moulines (2013), 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 Défossez and Bach (2015); Dieuleveut and Bach (2016). See also (Lan, 2012). Analysis has been extended to "tail averaging" (Jain et al., 2016), 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 proposed to improve the speed at which initial conditions were forgotten, for example using acceleration when the noise is additive (Dieuleveut et al., 2016; Jain et al., 2017).

Link between discretization of ergodic diffusions and SGD. 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 Talay and Tubaro (1990) and Mattingly et al. (2002) in the case of the Euler-Maruyama discretization. Then Talay and Tubaro (1990) 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 Abdulle et al. (2014) and Chen et al. (2015). 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 similarly to the Euler-Maruyama discretization, SGD has a weak error of order γ. Finally, Durmus et al. (2016) proposed and analyzed the use of Richardson-Romberg extrapolation applied to the stochastic gradient Langevin dynamics (SGLD) algorithm. This methods introduced by Welling and Teh (2011) combines SGD and the Euler-Maruyama discretization of the Langevin diffusion associated to a target probability measure. 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.

2.3 Summary and discussion of main results

Under the stated assumptions, the Markov chain $(\theta_k^{(\gamma)})_{k \geq 0}$ admits a unique invariant/stationary distribution π_γ which admits a moment of order 2, see Theorem 3 in Section 3. Recall that π_γ is a stationary distribution of this Markov chain if, when $\theta_0^{(\gamma)}$ is distributed according to π_γ, then $\theta_1^{(\gamma)}$ is distributed according to π_γ as well. In the next section, by two different methods (Theorem 2 and Theorem 5), we show that under suitable conditions on f and the noise $(\varepsilon_k)_{k \geq 1}$ that there exists $C \geq 0$ such that for all $\gamma \geq 0$, small enough

$$\bar{\theta}_\gamma = \int_{\mathbb{R}^d} \vartheta \pi_\gamma(\mathrm{d}\vartheta) = \theta_* + C\gamma + O(\gamma^2) \ .$$

Using Theorem 2, we get that for γ small enough and all $k \geq 1$,

$$\mathbb{E}(\bar{\theta}_k^{(\gamma)} - \theta_*) = \frac{A(\theta_0, \gamma)}{k} + C\gamma + O(\gamma^2) + O(e^{-k\mu\gamma}) \ . \tag{5}$$

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 (5) shows that $(2\bar{\theta}_k^{(\gamma)} - \bar{\theta}_k^{(2\gamma)})_{k\geq 0}$ satisfies

$$\mathbb{E}(2\bar{\theta}_k^{(\gamma)} - \bar{\theta}_k^{(2\gamma)} - \theta_*) = \frac{A(\theta_0,\gamma) - A(\theta_0,2\gamma)}{k} + O(\gamma^2) + O(e^{-k\mu\gamma}) \, ,$$

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 unaveraged gradient iterate $\theta_k^{(\gamma)}$ saturates rapidly, $\bar{\theta}_k^{(\gamma)}$ may already perform well enough to avoid saturation on real data-sets (Bach and Moulines, 2013). 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 the central limit theorem for a fixed $\gamma > 0$ and k goes to $+\infty$ for $(\bar{\theta}_k^{(\gamma)})_{k\geq 0}$, *i.e.* under appropriate conditions, there exist $B_1(\gamma)$ and $B_2(\gamma)$ such that

$$\mathbb{E}\left[\left\|\bar{\theta}_k^{(\gamma)} - \bar{\theta}_\gamma\right\|^2\right] = B_1(\gamma)/k + B_2(\gamma)/k^2 \, . \tag{6}$$

Combining (5) and (6) characterizes the bias/variance trade-off of SGD used to estimate θ_*.

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. The convergence is quantified with the Wasserstein distance (see e.g., Chapter 6 in Villani, 2009).

Limit distribution. A fundamental tool in Markov chain theory is the *Markov kernel*, which is the equivalent for continuous spaces of the *transition matrix* in finite state spaces. Let R_γ be the Markov kernel on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ associated with the SGD iterates $(\theta_k^{(\gamma)})_{k\geq 0}$, where $\mathcal{B}(\mathbb{R}^d)$ is the Borel σ-field of $\mathbb{R}^d$. We refer to (Meyn and Tweedie, 2009) for an introduction to Markov chain theory. For all initial distributions ν_0 on $\mathcal{B}(\mathbb{R}^d)$ and $k \in \mathbb{N}$, $\nu_0 R_\gamma^k$ denotes the law of $\theta_k^{(\gamma)}$ starting at θ_0 distributed according to ν_0. For any measure π on $\mathcal{B}(\mathbb{R}^d)$ and any measurable function $h : \mathbb{R}^d \to \mathbb{R}$, $\pi(h)$ denotes $\int h(\theta) d\pi(\theta)$ when it exists. Finally, for all $\theta \in \mathbb{R}^d$ and measurable function $h : \mathbb{R}^d \to \mathbb{R}$, $k \geq 1$, set $R_\gamma^k(\theta, \cdot) = \delta_\theta R_\gamma^k$ the distribution of $\theta_k^{(\gamma)}$ starting at θ and $R_\gamma^k h(\theta) = \int_{\mathbb{R}^d} h(\vartheta) \left\{\delta_\theta R_\gamma^k\right\}(\mathrm{d}\vartheta)$.

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 introduce the Wasserstein distance. For all probability measures ν and λ on $\mathcal{B}(\mathbb{R}^d)$, such that $\int_{\mathbb{R}^d} \|\theta\|^2 \, \mathrm{d}\nu(\theta) < +\infty$ and $\int_{\mathbb{R}^d} \|\theta\|^2 \, \mathrm{d}\lambda(\theta) \leq +\infty$, 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 $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $\xi(\mathsf{A} \times \mathbb{R}^d) = \nu(\mathsf{A})$, $\xi(\mathbb{R}^d \times \mathsf{A}) = \lambda(\mathsf{A})$.

Proposition 1. *Assume **A**1-**A**2-**A**3-**A**4(2), for any step size $\gamma < L^{-1}$, the Markov chain $(\theta_k^{(\gamma)})_{k\geq 0}$ defined by the recursion* (1)*, admits a unique stationary distribution π_γ such that $\int_{\mathbb{R}^d} \|\vartheta\|^2 \, \mathrm{d}\pi_\gamma(\vartheta) < +\infty$. In addition for all $\theta \in \mathbb{R}^d$, $k \in \mathbb{N}$:*

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

Proof. The proof is postponed to Appendix B.1. □

To prove the existence of the limit, one shows that for any x, $(R_\gamma^k(x,\cdot))_{k\geq 0}$ is a Cauchy sequence in a particular Polish space. We can thus define a point-wise limit, and show that it is unique. This uses the strong convexity, smoothness and the Lipschitzness of the noise.

As a consequence of Proposition 1, the expectation of $\bar{\theta}_k^{(\gamma)} = \frac{1}{k+1} \sum_{i=0}^k \theta_i^{(\gamma)}$ converges $\int_{\mathbb{R}^d} \vartheta \mathrm{d}\pi_\gamma(\vartheta)$ as k goes to infinity at a rate of order $O(k^{-1})$, see Theorem 12 in Appendix C.

3.1 Expansion of moments under π_γ when γ is in a neighborhood of 0

In this paragraph, 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 $\bar{\theta}_\gamma = \theta_* + O(\gamma)$. By simple developments of Equation (1) at the equilibrium, we propose expansions of the first two moments of the chain. It extends (Pflug, 1986; Ljung et al., 1992) 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_Σ is a quadratic function, *i.e.*, f' is affine, since π_γ is invariant for $(\theta_k^{(\gamma)})_{k\geq 0}$ then if $\theta_0^{(\gamma)}$ is distributed according to π_γ, since $\theta_1^{(\gamma)}$ is distributed according to π_γ as well and $\theta_1^{(\gamma)} = \theta_0^{(\gamma)} - \gamma f'(\theta_0^{(\gamma)}) + \gamma\varepsilon_1(\theta_0^{(\gamma)})$ taking expectations on both sides, we get $\int_{\mathbb{R}^d} f'(\vartheta)\mathrm{d}\pi_\gamma(\vartheta) = 0$ which, by linearity of f' imposes that $f'(\bar{\theta}_\gamma) = 0$ and thus that $\bar{\theta}_\gamma = \theta_*$. This implies that the averaged iterate converges to θ_*, see e.g. Bach and Moulines (2013). Moreover, as shown in Appendix B.3, we can also compute exactly the second moment as $\int_{\mathbb{R}^d}(\theta-\theta_*)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta) = \gamma(\Sigma\otimes I + I\otimes\Sigma - \gamma\Sigma\otimes\Sigma)^{-1}\int_{\mathbb{R}^d}\varepsilon_1(\theta)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta)$, where we denote, for any $\theta\in\mathbb{R}^d$, $\theta^{\otimes 2} := \theta\theta^\top$, where for any matrices $M, N \in \mathbb{R}^{d\times d}$, $M\otimes N$ is defined as the following operator from $\mathbb{R}^{d\times d}$ into $\mathbb{R}^{d\times d}$ such that $M\otimes N : P \mapsto MPN$.

General case. While the quadratic case led to particularly simple exact expressions, in general, we can only get a first order development of these expectations as $\gamma \to 0$ (proofs are given in Appendix B.3). Note that it improved on (Pflug, 1986), which shows a similar expansion but an error of order of $O(\gamma^{3/2})$.

Theorem 2 (Properties under stationarity, general case)**.** *Let $\gamma < 1/L$ and assume **A**1-**A**2-**A**3-**A**4(4). Then*

$$\bar{\theta}_\gamma - \theta_* = \gamma f''(\theta_*)^{-1} f'''(\theta_*)\Big(\big[f''(\theta_*)\otimes I + I\otimes f''(\theta_*)\big]^{-1}\int_{\mathbb{R}^d}\varepsilon(\theta)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta)\Big) + O(\gamma^2)$$

$$\int_{\mathbb{R}^d}(\theta-\theta_*)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta) = \gamma\big[f''(\theta_*)\otimes I + I\otimes f''(\theta_*)\big]^{-1}\int_{\mathbb{R}^d}\varepsilon(\theta)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta) + O(\gamma^2)\ ,$$

where π_γ is the stationary distribution of the Markov chain $(\theta_k^{(\gamma)})_{k\geq 0}$ defined by the recursion (1) *and $\bar{\theta}_\gamma$ is given by* (3)*.*

Proof. The proof is postponed to Appendix B.3. □

This shows that $\gamma \mapsto \bar{\theta}_\gamma$ is a differentiable function at $\gamma = 0$. The "drift" $\bar{\theta}_\gamma - \theta_*$ can be understood as an additional error occurring because the function is non quadratic and the step sizes are not decaying to zero. The mean under the limit distribution is at distance γ from θ_* while the final iterate oscillates in a sphere of radius proportional to $\sqrt{\gamma}$, as $\int_{\mathbb{R}^d}\|\theta-\theta_*\|\pi_\gamma(\mathrm{d}\theta) \leq \sqrt{\gamma}\,\mathrm{tr}^{1/2}(\big[f''(\theta_*)\otimes I + I\otimes f''(\theta_*)\big]^{-1}\int_{\mathbb{R}^d}\varepsilon(\theta)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta))$, where for any matrix $M\in R^{d\times d}$, $\mathrm{tr}(M)$ is the trace of M, *i.e.*, the sum of diagonal elements of the matrix M.

3.2 Expansion for a given $\gamma > 0$ when k tends to $+\infty$

In this Section, we analyze the convergence of $\bar{\theta}_k^{(\gamma)}$ to $\bar{\theta}_\gamma$, when $k\to\infty$, and the convergence of $\mathbb{E}\left[\left\|\bar{\theta}_k^{(\gamma)} - \bar{\theta}_\gamma\right\|^2\right]$ to 0. Under suitable conditions (Meyn and Tweedie, 1993; Jones, 2004), $\bar{\theta}_k^{(\gamma)}$ satisfies a central limit theorem: $\sqrt{k}\left(\bar{\theta}_k^{(\gamma)} - \bar{\theta}_\gamma\right) \xrightarrow{d} \mathcal{N}(0,\sigma_\varphi^2)$, where $\sigma_\varphi^2 \geq 0$. However, this result is purely asymptotic; we propose a new tighter development that describes how the initial conditions are forgotten: we prove 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 expressed easily as moments of solutions to several *Poisson equations*.

Poisson equation. For any Lipschitz function $\varphi : \mathbb{R}^d \to \mathbb{R}$, the convergence speed of $k^{-1}\sum_{i=0}^{k-1}\varphi(\theta_i^{(\gamma)})$ towards $\int_{\mathbb{R}^d}\varphi(\vartheta)\mathrm{d}\pi_\gamma(\vartheta)$ can be decomposed as a sum of two main terms, that can be expressed as moments of two Poisson solutions associated with φ which we now described. It shows in Appendix B.2 that the sequence of function $\{\theta \mapsto \sum_{i=1}^{k} R_\gamma^i\phi(\theta) - \pi_\gamma(\phi)\}_{k\geq 0}$ converges uniformly on all compact sets of $\mathbb{R}^d$. Define then $\psi_\gamma = \sum_{i=0}^{+\infty}\{R_\gamma^i\phi - \pi_\gamma(\phi)\}$. Note that ψ_γ satisfies $\pi_\gamma(\psi_\gamma) = 0$, $(I - R_\gamma)\psi_\gamma = \varphi$ and is Lipschitz, see Appendix B.2. ψ_γ will be referred to as the Poisson solution associated with φ.

For the convergence of $\bar{\theta}_k^{(\gamma)}$ to $\bar{\theta}_\gamma$, we thus introduce ψ_γ, the Poisson solution associated to $\varphi : \theta \mapsto \theta - \theta_*$, χ_γ^1 the Poisson solution associated to $\theta \mapsto \psi_\gamma(\theta)\psi_\gamma^\top(\theta)$, and finally χ_γ^2 the Poisson solution associated to $\theta \mapsto ((\psi_\gamma - \varphi)(\theta))^{\otimes 2}$. We then have:

Theorem 3 (Convergence of the Markov chain)**.** *Let $\gamma \in (0, 1/(2L))$ and assume **A**1-**A**2-**A**3-**A**4(4). Then for any starting point $\theta_0 \in \mathbb{R}^d$, setting $\rho := (1-\gamma\mu)^{1/2}$:*

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

where $(\bar{\theta}_k^{(\gamma)})_{k\geq 0}$ is given by (2) *and π_γ is its unique stationary distribution of the Markov chain defined by the recursion* (1)*.*

Proof. This result is a consequence of Theorem 9, proved in Appendix B.4.2. □

This bound for the second order moment decomposes as a sum of two terms: (i) a variance term, that scales as $1/k$, and does not depend on the initial distribution (but only on the asymptotic distribution π_γ), and (ii) a bias term, which scales as $1/k^2$, and depends on the initial distribution ν_0. The proof of this result relies on the following two identities, which illustrate that the associated Poisson solutions are introduced, $\mathbb{E}\left[\bar{\theta}_k^{(\gamma)}\right] - \theta_* = \frac{1}{k}\sum_{i=0}^{k-1}(R_\gamma^i\varphi)(\theta_0) = \pi_\gamma\varphi + \frac{1}{k}\psi_\gamma(\theta_0) + R_\gamma^k\psi_\gamma(\theta_0)$, using $R_\gamma^i\pi_\gamma(\varphi) = \pi_\gamma\varphi$, and $\sum_{i=0}^{k-1} R_\gamma^i(\varphi - \pi_\gamma(\varphi)) = \sum_{i=0}^{\infty} R_\gamma^i(\varphi - \pi_\gamma(\varphi)) - R_\gamma^k\sum_{i=0}^{\infty} R_\gamma^i(\varphi - \pi_\gamma(\varphi)) = \psi_\gamma - R_\gamma^k\psi_\gamma$. Finally, we have that $R_\gamma^k\psi_\gamma(\theta_0)$ converges to 0 at linear speed, using Proposition 1.

This result gives an exact closed form for the asymptotic bias and variance, for a fixed γ, and 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 $\gamma \to 0$. Indeed, part of the difficulty comes from the fact that as γ goes to zero, the Markov chain does not mix fast enough.

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 Défossez and Bach (2015).

Corollary 4. *Assume that f is a quadratic function f_Σ, **A**3 and **A**4(4). Consider the least mean squares algorithm iterates $(\theta_k^{(\gamma)})_{k\geq 0}$ starting from $\theta_0 \in \mathbb{R}^d$ with $\gamma L \leq 1/2$. Then*

$$\begin{aligned}
\mathbb{E}\left[(\bar{\theta}_k^{(\gamma)} - \theta_*)^{\otimes 2}\right] &= \frac{1}{k^2\gamma^2}\Sigma^{-1}\Omega(\theta_0 - \theta_*)^{\otimes 2}\Sigma^{-1} + \frac{1}{k}\Sigma^{-1}\left[\mathbb{E}_{\theta\sim\pi_\gamma}\left[\varepsilon_1^{\otimes 2}(\theta)\right]\right]\Sigma^{-1} \\
&\quad - \frac{1}{k^2\gamma}\Sigma^{-1}\Omega\left[\Sigma\otimes I + I\otimes\Sigma - \gamma T\right]^{-1}\left[\mathbb{E}\xi_1^{\otimes 2}\right]\Sigma^{-1} + O(\rho^k)\ ,
\end{aligned}$$

where $\rho = (1-\gamma\mu)^{1/2}$, $\Omega := (\Sigma\otimes I + I\otimes\Sigma - \gamma\Sigma\otimes\Sigma)(\Sigma\otimes I + I\otimes\Sigma - \gamma T)^{-1}$, $T : A \mapsto \mathbb{E}\left[(x^\top Ax)xx^\top\right]$ and ξ_1 is given by (4)*.*

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, with now $t \in \mathbb{R}$:

$$\dot{\theta}_t = -f'(\theta_t)\ . \tag{7}$$

Note that since $f'(\theta_*) = 0$ by definition of θ_* and **A**1, then θ_* is an equilibrium point of (7), *i.e.* $\theta_t = \theta_*$ for all $t \geq 0$ if $\theta_0 = \theta_*$. Under **A**2, (7) admits a unique solution on $\mathbb{R}_+$ for any starting point $\theta \in \mathbb{R}^d$. Denote by $(\phi_t)_{t\geq 0}$ the flow of (7), defined for all $\theta \in \mathbb{R}^d$ by $(\phi_t(\theta))_{t\geq 0}$ as the solution of (7) starting at θ.

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

$$D(\mathcal{A}) = \left\{ h : \mathbb{R}^d \to \mathbb{R} : \text{ for all } \theta \in \mathbb{R}^d, \lim_{t\to+\infty} \frac{h(\phi_t(\theta)) - h(\theta)}{t} \text{ exists } \right\}$$
$$\mathcal{A}h(\theta) = \lim_{t\to+\infty} t^{-1} \left\{ h(\phi_t(\theta)) - h(\theta) \right\} \text{ for all } h \in D(\mathcal{A}) , \quad \theta \in \mathbb{R}^d . \tag{8}$$

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

Under **A**1 and **A**2, $k \in \mathbb{N}$, $k \geq 1$, for any function $g : \mathbb{R}^d \to \mathbb{R}$ (extension to a function $g : \mathbb{R}^d \to \mathbb{R}^q$ can easily be done considering all assumptions and results coordinatewise), locally Lipschitz, denote by h_g the solution of the continuous Poisson equation defined for all $\theta \in \mathbb{R}^d$ by $h_g(\theta) = \int_0^\infty (g(\phi_s(\theta)) - g(\theta_*))ds$. Note that h_g is well-defined by Lemma 13-b) in Appendix D, since g is assumed to be locally Lipschitz. Note that by (8), we have for all $g : \mathbb{R}^d \to \mathbb{R}$, locally Lipschitz,

$$\mathcal{A}h_g(\theta) = g(\theta) - g(\theta_*) . \tag{9}$$

Under regularity assumptions on g (see Theorem 15), 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 implying the derivatives of h_g. For $g : \mathbb{R}^d \to \mathbb{R}$ and $k_1, k_2 \in \mathbb{N}$, $k_1 \geq 1$ consider the following assumptions.

A 5 (k_1, k_2)**.** *There exist $a_g, b_g \in \mathbb{R}_+$ such that $g \in C^{k_1}(\mathbb{R}^d)$ and for all $x \in \mathbb{R}^d$ and $i \in \{1, \cdots, k_1\}$, $\sup_{x\in\mathbb{R}^d} \left\| D^i g(\theta) \right\| \leq a_g \left\{ \|\theta - \theta_*\|^{k_2} + b_g \right\}$, where $D^i g$ is the differential of order i of g.*

A 6. *The functions $(\varepsilon_k)_{k\in\mathbb{N}^*}$ are i.i.d., and that the function $\mathcal{C}(\theta) : \theta \mapsto \mathbb{E}\left[\varepsilon_1(\theta)^{\otimes 2}\right]$ is three time continuously differentiable and there exists $M_\varepsilon \geq 0$ such that for all $\theta \in \mathbb{R}^d$, $\left\| D^i \mathcal{C}(\theta) \right\| \leq M_\varepsilon \left\{ 1 + \|\theta - \theta_*\|^{k_\varepsilon} \right\}$ for $i \in \{1, 2, 3\}$.*

Theorem 5. *Assume **A**1-**A**2-**A**3-**A**4$(2(q+3))$-**A**6, for $q \in \mathbb{N}$. Let $g : \mathbb{R}^d \to \mathbb{R}$ satisfying **A**5$(5, q)$. Then there exists $C_{2(q+3)}$ only depending on q such that for all $\gamma \in \left(0, C_{2(q+3)}\right)$, $k \in \mathbb{N}^*$ and $\theta_0 \in \mathbb{R}^d$ such that*

$$\mathbb{E}\left[k^{-1} \sum_{i=1}^k \left\{ g(\theta_i^{(\gamma)}) - g(\theta_*) \right\} \right] = \frac{\mathbb{E}\left[h_g(\theta_{k+1}^{(\gamma)}) \right] - h_g(\theta_0)}{k\gamma}$$
$$- (\gamma/2) h_g''(\theta_*) \mathbb{E}\left[\{\varepsilon(\theta_*)\}^{\otimes 2} \right] + \frac{\gamma}{k} A_1(\theta_0) + \gamma^2 A_2(\theta_0, k) , \tag{10}$$

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

$$A_1(\theta_0) \leq C \left\{ 1 + \|\theta_0 - \theta_*\|^{q+2} \right\} , \; A_2(\theta_0, k) \leq C \left\{ 1 + \|\theta_0 - \theta_*\|^{q+3} / k \right\} .$$

Proof. The proof is postponed to Appendix E. □

First in the case where f' is linear, choosing for g the identity function, then $h_{\mathrm{Id}} = \int_0^{+\infty} \{\phi_s - \theta_*\} \mathrm{d}s = \Sigma^{-1}$, and we get that the first term in (10) vanishes which is natural since in that case $\bar{\theta}_\gamma = \theta_*$. Second by Lemma 14-c), we recover the first expansion of Theorem 2 for arbitrary objective functions f. Finally note that for all $q \in \mathbb{N}$, under appropriate conditions Theorem 5 implies that there exists $C_1, C_2(\theta_0) \geq 0$ such that $\mathbb{E}\left[k^{-1} \sum_{i=1}^k \|\theta_i^{(\gamma)} - \theta_*\|^{2q} \right] = C_1\gamma + C_2(\theta_0)/n + O(\gamma^2)$.

4 Experiments

We performed experiments on simulated data, for logistic regression, with $n = 10^7$ observations, for $d = 10$ and 25. Results are presented in Figure 2. 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 (as $\|\theta_k^{(\gamma)} - \theta_*\| = O(\sqrt{\gamma})$). Note that the ratio between the convergence limits of the two sequences is roughly 2 in the un-averaged case, and 4 in the averaged case, which confirms the predicted limits. We consider Richardson Romberg iterates, which saturate at a much lower level, and performs much better than decaying step sizes (as $1/\sqrt{n}$) on the first iterations, as it forgets the initial conditions faster. Finally, we run the online-Newton (Bach and Moulines, 2013), which performs very well but has no convergence guarantee. On the Right plot, we also propose an estimator that uses 3 different step sizes to perform a higher order interpolation. More precisely, 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 that it does not saturate.

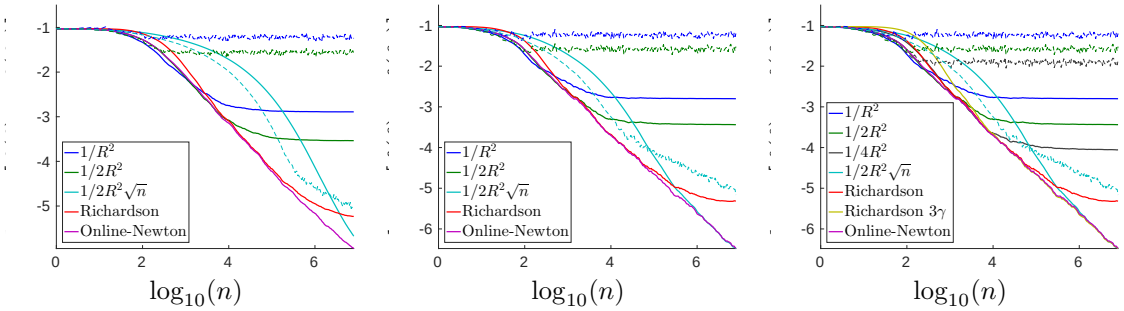

Figure 2: Synthetic data, logarithmic scales. Upper-left: logistic regression, $d = 12$, with averaged SGD with step-size $1/R^2$, $1/2R^2$, decaying step sizes as $1/2R^2\sqrt{n}$ (averaged (plain) and non-averaged (dashed)), Richardson Romberg extrapolated iterates, and online Newton iterates. Upper-right: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

5 Conclusion

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

6 Acknowledgments

The authors would like to thank Éric Moulines for helpful discussions. We acknowledge support from the chaire Economie des nouvelles donnees with the data science joint research initiative with the fonds AXA pour la recherche, and the Initiative de Recherche “Machine Learning for Large-Scale Insurance” from the Institut Louis Bachelier.

Notation

Denote by $\{\mathbf{e}_1, \dots, \mathbf{e}_d\}$ the canonical basis of $\mathbb{R}^d$. Let E and F be two vector spaces, 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. Let $n \in \mathbb{N}^*$, denote by $C^n(\mathbb{R}^d)$ the set of n times continuously differentiable functions from $\mathbb{R}^d$ to $\mathbb{R}$. Let $f \in C^n(\mathbb{R}^d)$, denote by $D^n f$ the n^{th} differential of f. Let $f \in C^1(\mathbb{R}^d)$, denote by ∇f the gradient of f. Let $f \in C^2(\mathbb{R}^d)$, denote by Δf the Laplacian of f. Denote by $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ the floor and ceiling function respectively. 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 $\mathcal{S}_{L,\mu}$ the set of μ-strongly convex and L-smooth functions on $\mathbb{R}^d$. By abuse of notation, we will denote sometimes $x^{\otimes 2} = xx^\top$.

In the next sections mainly devoted to proofs, we first introduce definitions and generalities about convex functions in Section A.1, then discuss extra different possible assumptions on the noise in Section A.3. We prove the existence of a limit distribution in Section B.1, and address asymptotic properties when $\gamma \to 0$ in Section A.1. We prove the convergence of the Markov chain in Section B.4, and study the relationship with the gradient flow in Section D.

A Generalities on convex and strongly convex functions

A.1 Definitions

Most of the following definitions can be found in Nesterov (2004). A continuously differentiable function f is **convex** if there exists for any $\theta, \eta \in \mathbb{R}^d$ we have:

$$f(\eta) \geq f(\theta) + \left\langle f'(\theta), \eta - \theta \right\rangle .$$

A continuously differentiable function f is **L-smooth** if its gradient is L-Lipschitz, i.e., if there exists a constant $L > 0$, such that for any $\theta, \eta \in \mathbb{R}^d$ we have:

$$\|f'(\eta) - f'(\theta)\| \leq L\|\eta - \theta\| .$$

A continuously differentiable function f is **μ-strongly convex** if there exists a constant $\mu > 0$, such that for any $\theta, \eta \in \mathbb{R}^d$ we have:

$$f(\eta) \geq f(\theta) + \left\langle f'(\theta), \eta - \theta \right\rangle + \frac{\mu}{2}\|\theta - \eta\|^2 .$$

Recall that θ_* refers to as $\arg\min_{\theta \in \mathbb{R}^d} f$, which is unique when f is strongly convex.

Let f be a L-smooth and μ-strongly convex function. Then for all $\theta, \eta \in \mathbb{R}^d$, it holds

$$\begin{aligned}
f(\theta) - f(\theta_*) &\geq \frac{\mu}{2}\|\theta - \theta_*\|^2 && (11)\\
f(\theta_n^{(\gamma)}) - f(\theta_*) &\leq L\|\theta_n^{(\gamma)} - \theta_*\|^2 && (12)\\
\left\langle f'(\theta) - f'(\eta), \theta - \eta \right\rangle &\geq \mu\|\theta - \eta\|^2 && (13)\\
\left\langle f'(\theta) - f'(\eta), \theta - \eta \right\rangle &\geq \frac{1}{L}\|f'(\theta) - f'(\eta)\|^2 && (14)\\
\left\langle f'(\theta) - f'(\eta), \theta - \eta \right\rangle &\geq \frac{L\mu}{L+\mu}\|\theta - \eta\|^2 + \frac{1}{L+\mu}\|f'(\theta) - f'(\eta)\|^2 . && (15)
\end{aligned}$$

The first two inequalities are direct consequences of the definition and the fact that $f'(\theta_*) = 0$. (13) is shown in (Nesterov, 2004, Chapter 2, (2.1.24)). (14) is the co-coercivity equation in (Zhu and Marcotte, 1996). (15) is a combination of the co-coercivity equation and of (13). It can be found in (Nesterov, 2004, Chapter 2, (2.1.24)),

A.2 Quadratic case

Consider the following assumption on f.

Q1. *There exists a positive definite matrix Σ such that $f = f_\Sigma := (\theta \mapsto \left\|\Sigma^{1/2}(\theta - \theta_*)\right\|^2)$.*

If there exists a positive definite matrix Σ such that $f = f_\Sigma := (\theta \mapsto \left\|\Sigma^{1/2}(\theta - \theta_*)\right\|^2)$, then **A**1 and **A**2 are satisfied, with μ the smallest eigenvalue of Σ, L its largest eigenvalue, and $M = 0$.

A.3 Discussion on assumptions on the noise

Assumption **A**4, made in the text, can be weakened in order to apply to settings where input observations are un-bounded (typically, Gaussian inputs would not satisfy Assumption **A**4). Especially, for most situations, we only need Assumption **A**7 below.

A 7. *(i) There exists* $\tau \geq 0$ *such that* $\{\mathbb{E}^{1/4}[\|\varepsilon_1(\theta_*)\|^4]\} \leq \tau$.

(ii) For all $\theta_1, \theta_2 \in \mathbb{R}^d$*, there exists* $L \geq 0$ *such that, for* $p = 2, \ldots, 4$*,*

$$\mathbb{E}\,\|f_1'(\theta_1) - f_1'(\theta_2)\|^p \leq L^{p-1}\,\|\theta_1 - \theta_2\|^{p-2}\left\langle \theta_1 - \theta_2, f'(\theta_1) - f'(\theta_2)\right\rangle , \tag{16}$$

We can also make the stronger assumption that the noise is independent of θ (referred to as the "semi-stochastic" setting, see Dieuleveut et al. (2016)), or more generally that the noise has a uniformly bounded fourth order moment.

A 8. *There exists* $\tau \geq 0$ *such that* $\sup_{\theta\in\mathbb{R}^d}\{\mathbb{E}^{1/4}[\|\varepsilon_1(\theta)\|^4]\} \leq \tau$.

Assumption **A**7 is the weakest, as it is satisfied for random design least mean squares and logistic regression with bounded fourth moment of the inputs. Note that we do not assume that gradient or gradient estimates are a.s. bounded, to avoid the need for a constraint on the space where iterates live. Of course Assumption **A**4 implies Assumption **A**7. Moreover, in the special case of Assumption **A**8 where the noise is independent of θ, then Assumption **A**4 is clearly satisfied under Assumption **A**2.

B Results on the Markov chain defined by SGD

B.1 Proof of Proposition 1

Let λ_1, λ_2 be two probability measures on $\mathcal{B}(\mathbb{R}^d)$ with finite second moment and $\gamma > 0$. Let $\theta_0^{(1)}, \theta_0^{(2)}$ be independent and distributed according to λ_1, λ_2 respectively, and $(\theta_k^{(1)})_{\geq 0}$,$(\theta_k^{(2)})_{k\geq 0}$ the SGD iterates associated with the step size γ, starting from $\theta_0^{(1)}$ and $\theta_0^{(2)}$ respectively and sharing the same noise, *i.e.* for all $k \geq 0$,

$$\begin{cases} \theta_{k+1}^{(1)} &= \theta_k^{(1)} - \gamma\big[f'(\theta_k^{(1)}) + \varepsilon_{k+1}(\theta_k^{(1)})\big] \\ \theta_{k+1}^{(2)} &= \theta_k^{(2)} - \gamma\big[f'(\theta_k^{(2)}) + \varepsilon_{k+1}(\theta_k^{(2)})\big] \ . \end{cases} \tag{17}$$

Therefore for all $k \geq 0$, the distribution of $(\theta_k^{(1)}, \theta_k^{(2)})$ belongs to $\Pi(\lambda_1 R_\gamma, \lambda_2 R_\gamma)$ defined in Section 3 in the main document. Then by definition of the Wasserstein distance,

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

using **A**3 for i), **A**7 for ii), and finally **A**1 for iii).

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

$$\begin{aligned} W_2^2(\lambda_1 R_\gamma^n, \lambda_2 R_\gamma^n) &\leq \mathbb{E}\left[\|\theta_n^{(1)} - \theta_n^{(2)}\|^2\right] \\ &\leq \rho\mathbb{E}\left[\|\theta_{n-1}^{(1)} - \theta_{n-1}^{(2)}\|^2\right] \leq \rho^n \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2 \,\mathrm{d}\lambda_1(x)\mathrm{d}\lambda_2(y) , \end{aligned} \tag{18}$$

By (Villani, 2009, Theorem 6.16), the space $\mathcal{P}_2(\mathbb{R}^d)$ of probability measures with second order moment on $\mathbb{R}^d$ endowed with W_2 is a Polish space. As a consequence of (18) for $\lambda_2 = \lambda_1 R_\gamma^p$, for $p \in \mathbb{N}$, and Picard fixed point theorem, $(\lambda_1 R_\gamma^n)_{n\geq 0}$ is a Cauchy sequence and converges to a limit $\pi_\gamma^{\lambda_1} \in \mathcal{P}_2(\mathbb{R}^d)$:

$$\lim_{n\to+\infty} W_2(\lambda_1 R_\gamma^n, \pi_\gamma^{\lambda_1}) = 0 \; . \tag{19}$$

In addition by the triangle inequality

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

Thus taking the limits as $n \to +\infty$, 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 $n \in \mathbb{N}$, $n \geq 1$,

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

Using (18) and (19), we get taking $n \to +\infty$, $W_2(\pi_\gamma R_\gamma, \pi_\gamma) = 0$ and $\pi_\gamma R_\gamma = \pi_\gamma$. The fact that π_γ is the unique stationary distribution can be shown by contradiction and using (18).

Thus finally for $\lambda_1 = \delta_\theta$, $\lambda_2 = \pi_\gamma$, using the invariance of π_γ and (18), we get:

$$W_2^2(R_\gamma^n(\theta, \cdot), \pi_\gamma) \quad \leq \quad (1 - 2\mu\gamma(1-\gamma L))^n \int \|\theta - \vartheta\|^2 \mathrm{d}\pi_\gamma(\vartheta) \; .$$

B.2 Existence of Poisson solutions

Using the process $(\theta_{k,\gamma}^{(1)})_{\geq 0}$,$(\theta_{k,\gamma}^{(2)})_{k\geq 0}$ defined by (17) with $\lambda_1 = \delta_\theta$ and $\lambda_2 = \pi_\gamma$ and (18), we have if h is L_h-Lipschitz, for any $x \in \mathbb{R}^d$, any $n \in \mathbb{N}^*$:

$$\begin{aligned} \left|R_\gamma^n(h - \pi_\gamma(h))(\theta)\right| \quad &\leq \quad L_h W_2^2(R_\gamma^n(\theta, \cdot), \pi_\gamma) \\ &\leq \quad L_h (1 - 2\mu\gamma(1-\gamma L))^{n/2} \left(\int \|\theta - \vartheta\|^2 \mathrm{d}\pi_\gamma(\vartheta) \right)^{1/2} \; . \end{aligned} \tag{20}$$

In addition, for any $(\theta, \vartheta) \in \mathbb{R}^d \times \mathbb{R}^d$, $n \in \mathbb{N}^*$, using (17):

$$\begin{aligned} \left\|R_\gamma^n h(\theta) - R_\gamma^n h(\vartheta)\right\| \quad &\leq \quad L_h W_2^2(R_\gamma^n(\theta, \cdot), R_\gamma^n(\vartheta, \cdot)) \\ &\leq \quad L_h (1 - 2\mu\gamma(1-\gamma L))^{n/2} \|\theta - \vartheta\| \; . \end{aligned} \tag{21}$$

As a consequence by (20), for any Lipschitz continuous function φ and any $\theta \in \mathbb{R}^d$, $\{\theta \mapsto \sum_{i=1}^k (R_\gamma^i \varphi(\theta) - \pi_\gamma(\varphi))\}_{k\geq 0}$ converges absolutely on all compact sets of $\mathbb{R}^d$. Denote by ψ_γ the limit associated with this sequence: $\psi_\gamma : \theta \mapsto \sum_{i=1}^\infty (R_\gamma^i \varphi(\theta) - \pi_\gamma(\varphi))$. By (21), ψ_γ is also Lipschitz continuous. This function is called the solution to the Poisson equation since it satisfies $(I - R_\gamma)\psi_\gamma = \varphi - \pi_\gamma(\phi)$. Moreover, $\pi_\gamma(\psi_\gamma) = 0$.

B.3 Asymptotic properties of the chain, behavior under equilibrium, and drift.

In the following, we consider the function $\varphi_1 : \theta \mapsto \theta - \theta_* \in \mathbb{R}^d$, and the function $\varphi_2 : \theta \mapsto (\theta - \theta_*)(\theta - \theta_*)^\top \in \mathbb{R}^{d\times d}$. In the quadratic case, we give an exact formula for the expectation under the limit distribution of these two terms. For the general case, we propose a first order development of these expectations.

The most important quantity, as we are eventually interested in the behavior of the averaged iterate $\bar{\theta}_n^{(\gamma)}$, is the expectation of the identity function under the limit distribution, $\bar{\theta}_\gamma$ defined by (3).

This part extends existing ideas from the literature to prove that $\gamma^{-1/2}(\pi_\gamma - \theta_*)$ converges in distribution to a normal law when $\gamma \to 0$. See for example (Pflug, 1986; Ljung et al., 1992). We consider the Markov chain under the limiting stationary distribution, together with a Taylor expansion of the function around the optimal point θ_*, in order to analyze how the average under the stationary distribution $\bar{\theta}_\gamma$ deviates from θ_*.

Analysis is carried through the dynamic (1) at stationarity, *i.e.* we assume that θ_0 is distributed according to π_γ given by the study of the equilibrium equation: under stationarity, *i.e.*, if $\theta_n^{(\gamma)} \sim \pi_\gamma$,

$$\theta_{n+1}^{(\gamma)} \stackrel{d}{=} \theta_n^{(\gamma)} - \gamma f'(\theta_n^{(\gamma)}) - \gamma\varepsilon_{n+1}(\theta_n^{(\gamma)}) \stackrel{d}{=} \pi_\gamma. \tag{22}$$

In order to get a first order development of $\bar{\theta}_\gamma$ around θ_*, we use the definition of the stationary distribution. We are going to use this equality several times to obtain information on θ's first moments under π_γ. The first consequence of this equation is that, taking expectations on both sides,

$$\int_{\mathbb{R}^d} f'(\theta)\pi_\gamma(\mathrm{d}\theta) \quad = \quad 0\,. \tag{23}$$

Lemma 6 (Properties under stationarity, Quadratic case)**.**
We consider, the stochastic gradient descent algorithm (1)*, for the quadratic function* $f_\Sigma(\theta) := \left\|\Sigma^{1/2}(\theta-\theta_*)\right\|^2$*. Then the mean value under the stationary distribution of the iterate is the optimal point:*

$$\begin{aligned}\bar{\theta}_\gamma &= \int_{\mathbb{R}^d} \theta\pi_\gamma(\mathrm{d}\theta) = \theta_* \\ \int_{\mathbb{R}^d} (\theta-\theta_*)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta) &= \gamma(\Sigma\otimes I + I\otimes\Sigma - \gamma\Sigma\otimes\Sigma)^{-1}\int_{\mathbb{R}^d}\varepsilon_1(\theta)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta)\,.\end{aligned}$$

Moreover, for the least mean squares algorithm, as defined described in the examples in Section 2.1,

$$\begin{aligned}\theta_n^{(\gamma)} - \theta_* &= (I-\gamma\Sigma)\left(\theta_{n-1}^{(\gamma)} - \theta_*\right) + \gamma\varepsilon_n(\theta_{n-1}^{(\gamma)}) \\ \varepsilon_n(\theta_{n-1}^{(\gamma)}) &= (\Sigma - x_n\otimes x_n)(\theta_{n-1}^{(\gamma)} - \theta_*) + (y_n - \langle\theta_*, x_n\rangle)x_n\,,\end{aligned}$$

we have another formula:

$$\int_{\mathbb{R}^d} (\theta-\theta_*)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta) \quad = \quad \gamma(\Sigma\otimes I + I\otimes\Sigma - \gamma M)^{-1}\mathbb{E}[\xi_1^{\otimes 2}]\,,$$

where in the last equation, M *is an operator on matrices such that* $M : A \mapsto \mathbb{E}[x_n x_n^\top A x_n x_n^\top]$*, and* $\xi_n = (y_n - \langle\theta_*, x_n\rangle)x_n$ *is the additive part of the noise (the part that does not depend on* θ*).*

Proof. The first part directly comes from Equation (23) and the fact that gradients of f_Σ are linear: $\int_{\mathbb{R}^d} f'(\theta)\pi_\gamma(\mathrm{d}\theta) = \Sigma\int_{\mathbb{R}^d}\theta-\theta_*\pi_\gamma(\mathrm{d}\theta) = 0$, thus $\int_{\mathbb{R}^d}\theta\pi_\gamma(\mathrm{d}\theta) = \theta_*$.

The second part comes from the development of Equation (22):

$$\begin{aligned}(\theta_1^{(\gamma)} - \theta_*)^{\otimes 2} &\stackrel{d}{=} \left((I-\gamma\Sigma)\left(\theta_0^{(\gamma)} - \theta_*\right) + \gamma\varepsilon_1(\theta_0^{(\gamma)})\right)^{\otimes 2} \\ \mathbb{E}(\theta_1^{(\gamma)} - \theta_*)^{\otimes 2} &= (I-\gamma\Sigma)\mathbb{E}\left(\theta_0^{(\gamma)} - \theta_*\right)^{\otimes 2}(I-\gamma\Sigma) + \gamma^2\mathbb{E}\left(\varepsilon_1(\theta_0^{(\gamma)})\right)^{\otimes 2} \\ \mathbb{E}(\theta_1^{(\gamma)} - \theta_*)^{\otimes 2} &= (I - \gamma\Sigma\otimes I - \gamma I\otimes\Sigma + \gamma^2\Sigma\otimes\Sigma)\mathbb{E}\left(\theta_0^{(\gamma)} - \theta_*\right)^{\otimes 2} \\ &\qquad + \gamma^2\mathbb{E}\left(\varepsilon_1(\theta_0^{(\gamma)})\right)^{\otimes 2}\,,\end{aligned} \tag{24}$$

Thus as if $\theta_0^{(\gamma)} \sim \pi_\gamma$, then $\theta_1^{(\gamma)} \sim \pi_\gamma$:

$$\int_{\mathbb{R}^d} (\theta-\theta_*)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta) \quad = \quad \gamma(\Sigma\otimes I + I\otimes\Sigma - \gamma\Sigma\otimes\Sigma)^{-1}\int_{\mathbb{R}^d}\varepsilon_1(\theta)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta)\,.$$

Similarly, starting from:

$$\theta_1^{(\gamma)} - \theta_* \quad = \quad (I-\gamma x_1\otimes x_1)\left(\theta_0^{(\gamma)} - \theta_*\right) + \gamma\xi_1\,,$$

using the fact that $\mathbb{E}[x_n x_n^\top] = \Sigma$ and the definition of M, one gets:

$$\int_{\mathbb{R}^d} (\theta-\theta_*)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta) \quad = \quad \gamma(\Sigma\otimes I + I\otimes\Sigma - \gamma M)^{-1}\mathbb{E}[\xi_1^{\otimes 2}]\,.$$

Which concludes the proof. □

Lemma 7. *Assume **A**1, **A**2, **A**3, **A**7. Then*

$$\mathbb{E}\left[-2\gamma\Big\langle f'_{n+1}(\theta_n^{(\gamma)}),\theta_n^{(\gamma)}-\theta_*\Big\rangle+\gamma^2\left\|f'_{n+1}(\theta_n^{(\gamma)})\right\|^2|\mathcal{F}_n\right]\leq -2\gamma\mu(1-\gamma L)\left\|\theta_n^{(\gamma)}-\theta_*\right\|^2+2\gamma^2\tau^2\ ,$$

where $f'_n=\varepsilon_n+f'$ for all $n\geq 1$ and $(\theta_n^{(\gamma)})_{n\geq 0}$ is given by (1).

Proof. Under **A**7, we have:

$$\begin{aligned}\mathbb{E}\left[\left\|f'_{n+1}(\theta_n^{(\gamma)})\right\|^2|\mathcal{F}_n\right] &\leq 2\left(\mathbb{E}\left[\left\|f'_{n+1}(\theta_n^{(\gamma)})-f'_{n+1}(\theta_*)\right\|^2\right]+\mathbb{E}\left[\left\|f'_{n+1}(\theta_*)\right\|^2|\mathcal{F}_n\right]\right)\\ &\leq 2\left(\mathbb{E}\left[\left\|f'_{n+1}(\theta_n^{(\gamma)})-f'_{n+1}(\theta_*)\right\|^2|\mathcal{F}_n\right]+\tau^2\right)\\ &\leq 2\left(L\mathbb{E}\left[\Big\langle f'_{n+1}(\theta_n^{(\gamma)})-f'_{n+1}(\theta_*),\theta_n^{(\gamma)}-\theta_*\Big\rangle|\mathcal{F}_n\right]+\tau^2\right)\\ &\leq 2\left(L\Big\langle f'(\theta_n^{(\gamma)})-f'(\theta_*),\theta_n^{(\gamma)}-\theta_*\Big\rangle+\tau^2\right)\ .\end{aligned}$$

Combining this result and **A**1 concludes the proof. □

Lemma 8 (Properties under stationarity, general case)**.**
*If f satisfies Assumptions **A**1, **A**2, and we study stochastic gradient descent under Assumptions **A**3, **A**7, we have:*

$$\bar{\theta}_\gamma-\theta_*=\frac{1}{2}\gamma f''(\theta_*)^{-1}f'''(\theta_*)\Big(\big[f''(\theta_*)\otimes I+I\otimes f''(\theta_*)\big]^{-1}\int_{\mathbb{R}^d}\varepsilon(\theta)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta)\Big)+O(\gamma^2)$$

$$\int_{\mathbb{R}^d}(\theta-\theta_*)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta) = \gamma\big[f''(\theta_*)\otimes I+I\otimes f''(\theta_*)\big]^{-1}\int_{\mathbb{R}^d}\varepsilon(\theta)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta)+O(\gamma^2)\ .$$

This lemma improves some result of (Pflug, 1986), and proves that the residual term is of order $O(\gamma^2)$ (we first prove that it is of order $O(\gamma^{3/2})$) and then improve on that result.

Proof. As before, the proof relies on the analysis of the recursion under stationarity. That is we consider $\theta_0^{(\gamma)}\sim\pi_\gamma$ (thus $\theta_1^{(\gamma)}\sim\pi_\gamma$), and expand the stochastic gradient recursion:

$$\begin{aligned}\theta_1^{(\gamma)} &= \theta_0^{(\gamma)}-\gamma f'_1(\theta_0^{(\gamma)})\\ &= \theta_0^{(\gamma)}-\gamma\left(f'(\theta_0^{(\gamma)})+\varepsilon_1(\theta_0^{(\gamma)})\right)\ .\end{aligned}$$

For simplicity, in the rest of the proof, we skip the explicit dependence in γ in $\theta_i^{(\gamma)}$, for $i\in\{0,1\}$. We only denote it θ_i.

We first prove that:

$$\bar{\theta}_\gamma-\theta_*=\frac{1}{2}\gamma f''(\theta_*)^{-1}f'''(\theta_*)\Big(\big[f''(\theta_*)\otimes I+I\otimes f''(\theta_*)\big]^{-1}\mathbb{E}\varepsilon^{\otimes 2}\Big)+O(\gamma^{3/2})\ .$$

We first notice that $\mathbb{E}_{\pi_\gamma}\|\theta-\theta_*\|=O(\gamma^{1/2})$, which will be used several times in the following. Indeed, if $\theta_0\sim\pi_\gamma$:

$$\begin{aligned}\mathbb{E}\left[\|\theta_1-\theta_*\|^2\right] &= \mathbb{E}\left[\|\theta_0-\theta_*-\gamma f'_1(\theta_0)\|^2\right]\\ &= \mathbb{E}\left[\|\theta_0-\theta_*\|^2-2\gamma\left\langle f'_1(\theta_0),\theta_0-\theta_*\right\rangle-\gamma^2\left\|f'_1(\theta_0)\right\|^2\right]\\ \Leftrightarrow 0 &\leq -2\gamma\mu\mathbb{E}\left[\|\theta_0-\theta_*\|^2\right]+\gamma^2\tau^2\end{aligned}$$

Using Lemma 7, under Assumption **A**7, with τ^2 the bound on $\mathbb{E}[\|\varepsilon_1(\theta_*)\|^2]$. Thus we have $\mathbb{E}_{\pi_\gamma}[\|\theta-\theta_*\|^2]\leq\frac{\gamma\tau^2}{2\mu}$, and by Jensen, $\mathbb{E}_{\pi_\gamma}[\|\theta-\theta_*\|]\leq\frac{\gamma^{1/2}\tau}{\sqrt{2\mu}}=O(\gamma^{1/2})$. More generally, we show in Appendix C, in Lemma 11, that $\mathbb{E}_{\pi_\gamma}[\|\theta-\theta_*\|^4]=O(\gamma^2)$, and thus $\mathbb{E}_{\pi_\gamma}[\|\theta-\theta_*\|^3]=O(\gamma^{3/2})$.

We now use the following expression for the SGD recursion:

$$\theta_1=\theta_0-\gamma\left(f'(\theta_0)+\varepsilon_1(\theta_0)\right)\ .$$

For simplicity, in the following, we may denote: $\varepsilon_1 = \varepsilon_1(\theta_0)$. By definition, we have $\bar{\theta}_\gamma = \mathbb{E}_{\pi_\gamma}\theta$, and as it has been seen before, $\mathbb{E}_{\pi_\gamma} f'(\theta) = 0$.

At it has been proved above, $\mathbb{E}_{\pi_\gamma}\|\theta - \theta_*\|^2 = O(\gamma)$, which also implies by Jensen's inequality that $\|\bar{\theta}_\gamma - \theta_*\|^2 = O(\gamma)$. Using a Taylor expansion, we have that:

$$f'(\theta) = f''(\theta_*)(\theta - \theta_*) + \frac{1}{2}f'''(\theta_*)(\theta - \theta_*)^{\otimes 2} + O(\|\theta - \theta_*\|^3).$$

Where $f''(\theta_*)$ is the Hessian matrix of f, and $f'''(\theta_*)$ a third order tensor that acts on the second order tensor $(\theta - \theta_*)^{\otimes 2}$: $f'''(\theta_*)(\theta - \theta_*)^{\otimes 2}$ is a vector in $\mathbb{R}^d$, such that for $k \in [1; d]$, $(f'''(\theta_*)(\theta - \theta_*)^{\otimes 2})_k = \sum_{i,j=1}^{n} \frac{\partial^3 f}{\partial\theta_i \partial\theta_j \partial\theta_k}(\theta - \theta_*)_i(\theta - \theta_*)_j$.

$$0 \quad = \quad \mathbb{E}_{\pi_\gamma}\Big[f''(\theta_*)(\theta - \theta_*) + \frac{1}{2}f'''(\theta_*)(\theta - \theta_*)^{\otimes 2}\Big] + O(\gamma^{3/2}),$$

using the fact that f is $\mathcal{C}^4$, with bounded 4^{-th} derivative, and $\mathbb{E}_{\pi_\gamma}[\|\theta - \theta_*\|^3] = O(\gamma^{3/2})$. This leads to

$$f''(\theta_*)(\bar{\theta}_\gamma - \theta_*) + \frac{1}{2}f'''(\theta_*)\big[\mathbb{E}_{\pi_\gamma}(\theta - \theta_*)^{\otimes 2}\big] = O(\gamma^{3/2}) \; . \tag{25}$$

Moreover, we have:

$$\begin{aligned} \theta_1 - \theta_* &= \theta_0 - \theta_* - \gamma\big[f''(\theta_*)(\theta_0 - \theta_*) + \varepsilon_1 + O(\|\theta_0 - \theta_*\|)\big] \\ &= (I - \gamma f''(\theta_*))(\theta_0 - \theta_*) - \gamma\varepsilon_1 + \gamma O(\|\theta_0 - \theta_*\|). \end{aligned}$$

Taking the second order moment of this equation, and using the fact that $\mathbb{E}_{\pi_\gamma}[\varepsilon_1(\theta_0 - \theta_*)^\top] = \mathbb{E}_{\pi_\gamma}[\mathbb{E}[\varepsilon_1(\theta_0 - \theta_*)^\top|\mathcal{F}_0]] = \mathbb{E}_{\pi_\gamma}[\mathbb{E}[\varepsilon_1|\mathcal{F}_0](\theta_0 - \theta_*)^\top] = 0$, we get:

$$\mathbb{E}_{\pi_\gamma}(\theta - \theta_*)^{\otimes 2} = (I - \gamma f''(\theta_*))\mathbb{E}_{\pi_\gamma}(\theta - \theta_*)^{\otimes 2}(I - \gamma f''(\theta_*)) + \gamma^2\mathbb{E}_{\pi_\gamma}[\varepsilon_1^{\otimes 2}] + O(\gamma^{5/2}).$$

This leads to:

$$\mathbb{E}_{\pi_\gamma}(\theta - \theta_*)^{\otimes 2} = \gamma\big[f''(\theta_*) \otimes I + I \otimes f''(\theta_*)\big]^{-1}\mathbb{E}_{\pi_\gamma}[\varepsilon_1^{\otimes 2}] + O(\gamma^{3/2}). \tag{26}$$

And combining Equation (25) and Equation (26), we get:

$$\bar{\theta}_\gamma - \theta_* = \frac{1}{2}\gamma f''(\theta_*)^{-1}f'''(\theta_*)\Big(\big[f''(\theta_*) \otimes I + I \otimes f''(\theta_*)\big]^{-1}\mathbb{E}_{\pi_\gamma}[\varepsilon_1^{\otimes 2}]\Big) + O(\gamma^{3/2}) \; .$$

The rest of the proof is devoted to showing that the residual term is of order $O(\gamma^2)$. At that point, we have also proved that $\mathbb{E}[\theta - \theta_*] = O(\gamma)$. To find the next term in the development, we develop further each of the terms. We introduce the 4^{-th} order tensor $f^{(4)} \in \mathbb{R}^{d\times d\times d\times d}$, which acts on $\mathbb{R}^{d\times d\times d}$ to give a vector of $\mathbb{R}^d$. Using the following Taylor expansion, with f assumed to be $\mathcal{C}^5$:

$$\begin{aligned} \theta_1 - \theta_* &= \theta_0 - \theta_* - \gamma\big[f''(\theta_*)(\theta_0 - \theta_*) + \frac{1}{2}f^{(3)}(\theta_*)(\theta_0 - \theta_*)^{\otimes 2} \\ &\quad + \frac{1}{6}f^{(4)}(\theta_*)(\theta_0 - \theta_*)^{\otimes 3} + \varepsilon_1 + O(\|\theta_0 - \theta_*\|^4)\big] \; . \end{aligned} \tag{27}$$

Thus if $\theta_0 \sim \pi_\gamma$:

$$\begin{aligned} \mathbb{E}_{\pi_\gamma}[\theta - \theta_*] &= \mathbb{E}_{\pi_\gamma}[\theta - \theta_*] - \mathbb{E}_{\pi_\gamma}\Big[\gamma\big[f''(\theta_*)(\theta - \theta_*) + \frac{1}{2}f^{(3)}(\theta_*)(\theta - \theta_*)(\theta - \theta_*)^\top \\ &\quad + \frac{1}{6}f^{(4)}(\theta_*)(\theta - \theta_*)^{\otimes 3} + \varepsilon_1\big]\Big] + \gamma O(\gamma^2) \\ f''(\theta_*)\mathbb{E}_{\pi_\gamma}[\theta - \theta_*] &= -\mathbb{E}_{\pi_\gamma}\left[\frac{1}{2}f^{(3)}(\theta_*)(\theta - \theta_*)^{\otimes 2} + \frac{1}{6}f^{(4)}(\theta_*)(\theta - \theta_*)^{\otimes 3} + \varepsilon_1\right] + O(\gamma^2) \\ f''(\theta_*)(\bar{\theta}_\gamma - \theta_*) &= -\frac{1}{2}f^{(3)}(\theta_*)\mathbb{E}_{\pi_\gamma}[(\theta - \theta_*)^{\otimes 2}] - \frac{1}{6}f^{(4)}(\theta_*)\mathbb{E}_{\pi_\gamma}[(\theta - \theta_*)^{\otimes 3}] + O(\gamma^2) \; . \end{aligned} \tag{28}$$

Using Assumption 3 (implying $\mathbb{E}[\varepsilon_1(\theta_0)] = 0$). To get the next term in the development, we need to

- Expand $\mathbb{E}_{\pi_\gamma}[\theta-\theta_*]^{\otimes 2} = \square\gamma + \triangle\gamma^2 + o(\gamma^2)$;
- Expand $\mathbb{E}_{\pi_\gamma}[(\theta-\theta_*)^{\otimes 3}] = \blacksquare\gamma^2 + o(\gamma^2)$.

First, we have, squaring Equation (27) and taking expectations:

$$\begin{aligned}\mathbb{E}[\theta_1-\theta_*]^{\otimes 2} &= \mathbb{E}\Big[(I-\gamma f''(\theta_*))\,(\theta_0-\theta_*) + \frac{\gamma}{2} f^{(3)}(\theta_*)(\theta_0-\theta_*)^{\otimes 2} + \gamma\varepsilon_1 \\ &\quad + O(\gamma\,\|\theta_0-\theta_*\|^3)\Big]^{\otimes 2} \\ &= \mathbb{E}[\theta_0-\theta_*]^{\otimes 2} - \gamma(I\otimes f''(\theta_*) + f''(\theta_*)\otimes I)\mathbb{E}\big[(\theta-\theta_*)^{\otimes 2}\big] + O(\gamma^3) \\ &\quad + \frac{\gamma}{2}\Big((\theta_0-\theta_*) f^{(3)}(\theta_*)(\theta_0-\theta_*)^{\otimes 2} + [(\theta_0-\theta_*) f^{(3)}(\theta_*)(\theta_0-\theta_*)^{\otimes 2}]^\top\Big) \\ &\quad + \gamma^2\mathbb{E}\varepsilon_1^{\otimes 2} + \gamma\mathbb{E}[(I-\gamma f''(\theta_*))\,(\theta_0-\theta_*)\varepsilon_1^\top]\,.\end{aligned}$$

Where we have used:

- $\gamma^2\mathbb{E}\big[(\theta-\theta_*)^{\otimes 2}\big] = O(\gamma^3)$.
- $\mathbb{E}[(I-\gamma f''(\theta_*))\,(\theta_0-\theta_*)\varepsilon_1^\top] = 0$ (Assumption 3 again).

Under $\theta_0 \overset{d}{=} \theta_1 \sim \pi_\gamma$, and simplifying by $\mathbb{E}_{\pi_\gamma}[\theta-\theta_*]^{\otimes 2}$ left and right and dividing by γ:

$$\begin{aligned}(I\otimes f''(\theta_*) + f''(\theta_*)\otimes I)\mathbb{E}_{\pi_\gamma}\big[(\theta-\theta_*)^{\otimes 2}\big] = O(\gamma^2) - \mathbb{E}\frac{1}{2} f^{(3)}(\theta_*)(\theta-\theta_*)^{\otimes 3} - \\ \mathbb{E}[\frac{1}{2} f^{(3)}(\theta_*)(\theta-\theta_*)^{\otimes 3}]^\top - \gamma\mathbb{E}\varepsilon_1^{\otimes 2}. \qquad (29)\end{aligned}$$

We now show that $\mathbb{E}_{\pi_\gamma}[(\theta-\theta_*)^{\otimes 3}] = O(\gamma^2)$. It can then be used in both (29) and (28), to prove that the next leading term is indeed or order $O(\gamma^2)$ and not $\gamma^{3/2}$. To compute $\mathbb{E}_{\pi_\gamma}[(\theta-\theta_*)^{\otimes 3}]$ we use the second order development again:

$$\begin{aligned}\theta_1-\theta_* &= \theta_0-\theta_* - \gamma\big[f''(\theta_*)(\theta_0-\theta_*) + \varepsilon_1 + O(\gamma)\big] \\ &= (I-\gamma f''(\theta_*))(\theta_0-\theta_*) - \gamma\varepsilon_1 + O(\gamma^2).\end{aligned}$$

$$\mathbb{E}_{\pi_\gamma}(\theta-\theta_*)^{\otimes 2} = (I-\gamma f''(\theta_*))\mathbb{E}_{\pi_\gamma}(\theta-\theta_*)^{\otimes 2}(I-\gamma f''(\theta_*)) + \gamma^2\mathbb{E}\varepsilon^{\otimes 2} + O(\gamma^{5/2}).$$

Let us denote in the following $\eta_i = \theta_i - \theta_*$, $i\in\{1,2\}$:

$$\begin{aligned}\mathbb{E}[\eta_1^{\otimes 3}] &= \mathbb{E}(\theta_1-\theta_*)^{\otimes 3} \\ &= \mathbb{E}\left((I-\gamma f''(\theta_*))\eta_0 - \gamma\varepsilon_1 + O(\gamma^2)\right)^{\otimes 3} \\ &= \mathbb{E}((I-(\gamma f''(\theta_*)\otimes I\otimes I + I\otimes\gamma f''(\theta_*)\otimes I + \ I\otimes I\otimes\gamma f''(\theta_*))(\eta_0)^{\otimes 3} \\ &\quad + O((\gamma^{2+3/2})) + \gamma^2\mathbb{E}[(I-\gamma f''(\theta_*))\eta_0\otimes\varepsilon_1^{\otimes 2} + \varepsilon_1\otimes(I-\gamma f''(\theta_*))\eta_0\otimes\varepsilon_1 \\ &\quad + \varepsilon_1^{\otimes 2}\otimes(I-\gamma f''(\theta_*))\eta_0] + \gamma^3\mathbb{E}[\varepsilon_1^{\otimes 3}] + 0 + O(\gamma^3)\,.\end{aligned}$$

Using the fact that $\mathbb{E}[\varepsilon_1]=0$, and the fact that $\mathbb{E}[O(\gamma^2)\otimes((I-\gamma f''(\theta_*))\eta)^{\otimes 2}] = O(\gamma^3)$ as $\mathbb{E}[\eta^{\otimes 2}] = O(\gamma)$. Thus, if $\theta_0 \overset{d}{=} \theta_1$, simplifying by $\mathbb{E}[\eta_i^{\otimes 3}]$:

$$\begin{aligned}\gamma\mathbf{M}\mathbb{E}[\eta_0^{\otimes 3}] &= \gamma^2\mathbb{E}\big[(I-\gamma f''(\theta_*))\eta_0\otimes\varepsilon_1^{\otimes 2} + \varepsilon_1\otimes(I-\gamma f''(\theta_*))\eta\otimes\varepsilon_1 \\ &\quad + \varepsilon_1^{\otimes 2}\otimes(I-\gamma f''(\theta_*))\eta_0\big] + \gamma^3\mathbb{E}[\varepsilon_1^{\otimes 3}] + 0 + O(\gamma^3)\,.\end{aligned}$$

With $\mathbf{M} = (f''(\theta_*)\otimes I\otimes I + I\otimes f''(\theta_*)\otimes I + \ I\otimes I\otimes f''(\theta_*)) : \mathbb{R}^{d\times d\times d}\to\mathbb{R}^{d\times d\times d}$. We need to bound the term $\mathbb{E}[(I-\gamma f''(\theta_*))\eta_0\otimes\varepsilon_1^{\otimes 2}]$ and its symmetric counterparts. We recall that ε_1 stands for $\varepsilon_1(\theta_0)$ and decompose it as the sum of an additive noise (independent on θ_0) and a multiplicative one: $\varepsilon_1(\theta_0) = \varepsilon_1(\theta_0) - \varepsilon_1(\theta_*) + \varepsilon_1(\theta_*)$. For the multiplicative part, under Assumption 7, $\mathbb{E}[\|\varepsilon_1(\theta_0)-\varepsilon_1(\theta_*)\|^2\,|\mathcal{F}_0] \le L[\|\theta_0-\theta_*\|^2]$, and thus $\mathbb{E}[(I-\gamma f''(\theta_*))\eta_0\otimes(\varepsilon_1(\theta_0)-\varepsilon_1(\theta_*))^{\otimes 2}] = O(\gamma^{3/2})$. For the additive part,

$$\begin{aligned}\mathbb{E}[(I-\gamma f''(\theta_*))\eta_0\otimes\varepsilon_1(\theta_*)^{\otimes 2}] &= \mathbb{E}[(I-\gamma f''(\theta_*))\eta_0\otimes\mathbb{E}[\varepsilon_1(\theta_*)^{\otimes 2}|\mathcal{F}_0]] \\ &= \mathbb{E}[(I-\gamma f''(\theta_*))\eta_0\otimes C] \\ &= (I-\gamma f''(\theta_*)(\bar\theta_\gamma-\theta_*)\otimes C\,,\end{aligned}$$

with $C = \mathbb{E}[\varepsilon_1(\theta_*)^{\otimes 2}] = \mathbb{E}[\varepsilon_1(\theta_*)^{\otimes 2}|\mathcal{F}_0]$ as $\varepsilon_1(\theta_*)^{\otimes 2}$ is independent of $\mathcal{F}_0$, and thus $\mathbb{E}[(I - \gamma f''(\theta_*))\eta_0 \otimes \varepsilon_1(\theta_*)^{\otimes 2}] = O(\gamma)$. Finally, for the crossed term, we use the fact that the multiplicative noise is Lipschitz to get the same result. Overall

$$\begin{aligned}\mathbf{M}\mathbb{E}_{\pi_\gamma}\left[(\theta-\theta_*)^{\otimes 3}\right] &= \gamma^2\left(\mathbb{E}_{\pi_\gamma}[\varepsilon_1^{\otimes 3}] + \frac{1}{\gamma}\mathbb{E}_{\pi_\gamma}[\eta_0\otimes\varepsilon_1^{\otimes 2} + \varepsilon_1\otimes\eta_0\otimes\varepsilon_1 + \varepsilon_1^{\otimes 2}\otimes\eta_0]\right) \\ &= O(\gamma^2) \end{aligned} \tag{30}$$

Combining (30) and the previously established results, we get the Lemma. □

B.4 Convergence of second order moments

B.4.1 Poisson equation

We now introduce the Poisson equation; for a function $\varphi : \mathbb{R}^d \to \mathbb{R}^q$ locally-Lipschitz, let $\psi : \mathbb{R}^d \to \mathbb{R}^q$ be a function such that $\pi_\gamma(\psi) = 0$ and the following equations:

$$(I - R_\gamma)\psi_f = \varphi - \pi_\gamma(\varphi) \tag{31}$$

$$\psi_f = \sum_{i=0}^{\infty} R_\gamma^i(\varphi - \pi_\gamma(\varphi)) \,, \tag{32}$$

such that for any $x \in \mathbb{R}^d$, $\psi_f(x) = \sum_{i=0}^{\infty} R_\gamma^i(\varphi - \pi_\gamma(\varphi))(x) = \sum_{i=0}^{\infty} \mathbb{E}\left[\varphi(\theta_i^{(\gamma)}(x))\right] - \pi_\gamma(\varphi)$. The convergence of this sum has already been proved for Lipschitz functions, using the contraction in Wasserstein distance between the law of iterates. More generally, for any locally Lipschitz function, Theorem 12, proved in Appendix C, shows that the solution to the Poisson equation exists, and is locally Lipschitz. As a consequence, we can consider recursively consider the solution to a Poisson equation associated to the solution of a Poisson equation.

B.4.2 Convergence theorem

Theorem 9. *Let $\varphi : \mathbb{R}^d \to \mathbb{R}^q$ be a locally Lipschitz function, let ψ be the solution of the Poisson Equation* (31)*. We assume that $\theta_0 \sim \nu_0$ for some initial distribution ν_0. We study Φ defined as the following random variable in $\mathbb{R}^q$.*

$$\Phi := \frac{1}{n}\sum_{i=0}^{n-1} \varphi(\theta_i^{(\gamma)}(\nu_0)) \,,$$

Then:

$$\mathbb{E}\Phi = \pi_\gamma(\varphi) + \frac{1}{n}\nu_0(\psi) + O(\rho^n) \,.$$

And if $\pi_\gamma(\varphi) = 0$:

$$\begin{aligned}\mathbb{E}(\Phi\Phi^\top) - (\mathbb{E}\Phi)(\mathbb{E}\Phi)^\top &= \frac{1}{n}\int_{\mathbb{R}^d}\left[\psi_\gamma(\theta)\psi_\gamma(\theta)^\top - (\psi_\gamma - \varphi)(\theta)(\psi_\gamma - \varphi)(\theta)^\top\right]\mathrm{d}\pi_\gamma(\theta) \\ &\quad + \frac{1}{n^2}\int_{\mathbb{R}^d}\left[\psi_\gamma(\theta)\psi_\gamma(\theta)^\top + \chi_\gamma^1(\theta) - \chi_\gamma^2(\theta)\right]\mathrm{d}\nu_0(\theta) + O(\rho^n) \,,\end{aligned}$$

where:

1. *$\rho := (1 - 2\mu\gamma(1 - \gamma L))^{1/2}$.*
2. *ψ_γ is the solution to the Poisson equation associated with φ.*
3. *χ_γ^1 is the solution to the Poisson equation associated with $\psi_\gamma\psi_\gamma^\top$.*
4. *χ_γ^2 is the solution to the Poisson equation associated with $(R_\gamma\psi_\gamma)(R_\gamma\psi_\gamma^\top)$.*

Proof. In the following proof, in order to improve readability, we skip the dependance on γ for $\theta_n^{(\gamma)}$, which is thus simply denoted θ_n. We have:

$$\begin{aligned}
\mathbb{E}\Phi &= \frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\varphi(\theta_i^{\nu_0})\right] = \frac{1}{n}\sum_{i=0}^{n-1}\nu_0(R_\gamma^n(\varphi)) \\
&= \pi_\gamma(\varphi) + \frac{1}{n}\sum_{i=0}^{n-1}\nu_0(R_\gamma^n(\varphi - \pi_\gamma(\varphi))) \\
&= \pi_\gamma(\varphi) + \frac{1}{n}\nu_0(\psi_\gamma) + \nu_0(R_\gamma^n(\psi_\gamma)) \\
&= \pi_\gamma(\varphi) + \frac{1}{n}\nu_0(\psi) + O(\rho^n)\ ,
\end{aligned}$$

with $\rho := (1-2\mu\gamma(1-\gamma L))^{1/2}$, and using the fact that $\nu_0(R_\gamma^n(\psi_\gamma)) = \nu_0(R_\gamma^n(\psi_\gamma - \pi(\psi_\gamma)))$. We now consider:

$$\begin{aligned}
\mathbb{E}\Phi\Phi^\top &= \frac{1}{n^2}\sum_{i,j=0}^{n-1}\mathbb{E}\varphi(\theta_i^{\nu_0})\varphi(\theta_j^{\nu_0})^\top \\
&= \frac{1}{n^2}\sum_{i=0}^{n-1}\left(\mathbb{E}\varphi(\theta_i^{\nu_0})\varphi(\theta_i^{\nu_0})^\top + \sum_{j=i+1}^{n-1}\left[\mathbb{E}\varphi(\theta_i^{\nu_0})\varphi(\theta_j^{\nu_0})^\top + \mathbb{E}\varphi(\theta_j^{\nu_0})\varphi(\theta_i^{\nu_0})^\top\right]\right) \\
&= -\frac{1}{n^2}\sum_{i=0}^{n-1}\nu_0(R_\gamma^i(\varphi(\cdot)\varphi(\cdot)^\top)) \\
&\quad + \frac{1}{n^2}\sum_{i=0}^{n-1}\left(\sum_{j=i+1}^{n-1}\left[\mathbb{E}\varphi(\theta_i^{\nu_0})\varphi(\theta_j^{\nu_0})^\top + \mathbb{E}\varphi(\theta_j^{\nu_0})\varphi(\theta_i^{\nu_0})^\top\right]\right) \\
&= -\frac{1}{n}\pi_\gamma(\varphi(\cdot)\varphi(\cdot)^\top) - \frac{1}{n^2}\nu_0\left(\sum_{i=0}^{\infty}R_\gamma^i\left((\varphi(\cdot)\varphi(\cdot)^\top) - \pi_\gamma(\varphi(\cdot)\varphi(\cdot)^\top)\right)\right) \\
&\quad + O(\rho^n) + \frac{1}{n^2}\sum_{i=0}^{n-1}\sum_{j=i}^{n-1}\left[\mathbb{E}\varphi(\theta_i^{\nu_0})(R_\gamma^{j-i}\varphi(\theta_i^{\nu_0}))^\top + \mathbb{E}(R_\gamma^{j-i}\varphi(\theta_i^{\nu_0}))\varphi(\theta_i^{\nu_0})^\top\right] \\
&= -\frac{1}{n}\pi_\gamma(\varphi(\cdot)\varphi(\cdot)^\top) - \frac{1}{n^2}\nu_0\left(\chi_\gamma^3\right) \\
&\quad + \frac{1}{n^2}\sum_{i=0}^{n-1}\left(\sum_{j=0}^{n-1-i}\left[\mathbb{E}\varphi(\theta_i^{\nu_0})(R_\gamma^j\varphi(\theta_i^{\nu_0}))^\top + \mathbb{E}(R_\gamma^j\varphi(\theta_i^{\nu_0}))\varphi(\theta_i^{\nu_0})^\top\right]\right).
\end{aligned}$$

With χ^3 the solution to the Poisson equation associated with $\varphi\varphi^\top$. Thus:

$$\begin{aligned}
\mathbb{E}\Phi\Phi^\top &= -\frac{1}{n}\pi_\gamma(\varphi(\cdot)\varphi(\cdot)^\top) - \frac{1}{n^2}\nu_0\left(\chi_\gamma^3\right) + O(\rho^n) \\
&\quad + \frac{1}{n^2}\sum_{i=0}^{n-1}\nu_0\left(R_\gamma^i\left[\varphi(\cdot)\psi_\gamma(\cdot) - \varphi(\cdot)R_\gamma^{n-i}\psi(\cdot)^\top\right] + \text{ symmetric term}\right)\ .
\end{aligned}$$

Using that $\frac{1}{n^2}\sum_{i=0}^{n-1}\nu_0\left(R_\gamma^i\left[\varphi(\cdot)R_\gamma^{n-i}\psi(\cdot)^\top\right]\right) = O(\rho^n)$, we get:

$$\begin{aligned}
\mathbb{E}\Phi\Phi^\top &= -\frac{1}{n}\pi_\gamma(\varphi(\cdot)\varphi(\cdot)^\top) - \frac{1}{n^2}\nu_0\left(\chi_\gamma^3\right) \\
&\quad + \frac{1}{n}\pi_\gamma\left(\varphi(\cdot)\psi_\gamma(\cdot)^\top\right) + \frac{1}{n^2}\nu_0(\chi_\gamma^4) \\
&\quad + \text{symmetric terms} + O(\rho^n)\ .
\end{aligned}$$

With χ^4 the solution to the Poisson equation associated with $\varphi\psi_\gamma^\top$.

For the first order terms, which scale as $\frac{1}{n}$, we have:

$$\begin{aligned}\mathbb{E}(\Phi\Phi^\top) - (\mathbb{E}\Phi)(\mathbb{E}\Phi)^\top &= \frac{1}{n}\pi_\gamma\left(-\varphi(\cdot)\varphi(\cdot)^\top + \varphi(\cdot)\psi_\gamma(\cdot)^\top + \psi_\gamma(\cdot)\varphi(\cdot)^\top\right) \\ &= \frac{1}{n}\pi_\gamma\left(-\varphi(\cdot)\varphi(\cdot)^\top + \varphi(\cdot)\psi(\cdot)^\top + \psi(\cdot)\varphi(\cdot)^\top\right) \\ &= \frac{1}{n}\pi_\gamma\left(-(\varphi-\psi)(\cdot)(\varphi-\psi)(\cdot)^\top + \psi(\cdot)\psi(\cdot)^\top\right) \\ &= \frac{1}{n}\pi_\gamma\left(-(R_\gamma\psi)(\cdot)(R_\gamma\psi)(\cdot)^\top + \psi(\cdot)\psi(\cdot)^\top\right) ,\end{aligned}$$

using the fact that for the solution to the Poisson equation: $\psi - R_\gamma\psi = \varphi$, *i.e.*, $\psi - \varphi = R_\gamma\psi$. This can also be written:

$$\mathbb{E}(\Phi\Phi^\top) - (\mathbb{E}\Phi)(\mathbb{E}\Phi)^\top = \frac{1}{n}\int_{\mathbb{R}^d}\left[\psi_\gamma(\theta)\psi_\gamma(\theta)^\top - (\psi_\gamma-\varphi)(\theta)(\psi_\gamma-\varphi)(\theta)^\top\right]\mathrm{d}\pi_\gamma(\theta) .$$

For the following order in $O(1/n^2)$, we have:

$$\begin{aligned}\mathbb{E}(\Phi\Phi^\top) - (\mathbb{E}\Phi)(\mathbb{E}\Phi)^\top - \frac{\text{term}}{n} &= \frac{-1}{n^2} + \frac{1}{n^2}\nu_0(-\chi_\gamma^3 + \chi_\gamma^4) + \text{ symmetric term} \\ &= \frac{1}{n^2}\nu_0(\chi_\gamma^1 - \chi_\gamma^2) ,\end{aligned}$$

using the linearity of R_γ and the fact that: $-\varphi\varphi^\top + \psi_\gamma\varphi^\top + \varphi\psi_\gamma^\top = -(\varphi-\psi)(\cdot)(\varphi-\psi)(\cdot)^\top + \psi(\cdot)\psi(\cdot)^\top$, thus: $\nu_0(-\chi_\gamma^3 + \chi_\gamma^4) = \nu_0(\chi_\gamma^1 - \chi_\gamma^2)$.
This is the expected result. □

B.4.3 Application in the quadratic case ($f = f_\Sigma$), for $\varphi = I$

We consider, the stochastic gradient descent algorithm (1), for the quadratic function $f_\Sigma(\theta) := \left\|\Sigma^{1/2}(\theta-\theta_*)\right\|^2$. We consider the classical stochastic approximation noise oracle of the least mean squares (LMS) algorithm:

$$\begin{aligned}\theta_{n,\gamma} - \theta_* &= (I - \gamma\Sigma)(\theta_{n-1,\gamma} - \theta_*) + \gamma\varepsilon_n(\theta_{n-1,\gamma}) \\ \varepsilon_n(\theta_{n-1,\gamma}) &= (\Sigma - x_n\otimes x_n)(\theta_{n-1,\gamma} - \theta_*) + (y_n - \langle\theta_*, x_n\rangle)x_n .\end{aligned}$$

We first recall the observation made in Appendix B.3: for quadratic functions, under the stationary distribution, the mean value of the iterate is the optimal point. According to Lemma 6, we have $\pi_\gamma(\varphi) = 0$. The following Lemma recovers result from Défossez and Bach (2015), as a corollary of our more general theorem.

Lemma 10. *If f is a quadratic function f_Σ, and we consider the LMS algorithm with $\gamma L \leq 1/2$, then with $\rho \leq (1-\gamma\mu)$, we have:*

$$\begin{aligned}\mathbb{E}\left[(\bar\theta_n^{(\gamma)} - \theta_*)^{\otimes 2}\right] &= \frac{1}{n^2\gamma^2}\Sigma^{-1}\Omega(\theta_0-\theta_*)^{\otimes 2}\Sigma^{-1} + \frac{1}{n}\Sigma^{-1}\left[\mathbb{E}_{\pi_\gamma}\varepsilon^{\otimes 2}\right]\Sigma^{-1} \\ &\quad -\frac{1}{n^2\gamma}\Sigma^{-1}\Omega\left[\Sigma\otimes I + I\otimes\Sigma - \gamma T\right]^{-1}\left[\mathbb{E}\xi^{\otimes 2}\right]\Sigma^{-1} .\end{aligned}$$

With $\Omega := (\Sigma\otimes I + I\otimes\Sigma - \gamma\Sigma\otimes\Sigma)(\Sigma\otimes I + I\otimes\Sigma - \gamma T)^{-1}$.

Moreover, the value of ρ is known: $\rho = (1 - 2\gamma\mu(1-\gamma L)) \leq (1-\gamma\mu)$ if $\gamma L \leq 1/2$, with $\mu = \lambda_{\min}(\Sigma)$.

Proof. We consider the linear function φ which is $\varphi(\theta) = \theta - \theta_*$. We then have that $\psi(\theta) = (\gamma\Sigma)^{-1}(\theta-\theta_*)$. Indeed from Equation (32), for any θ_0:

$$\psi(\theta_0) = \sum_{i=0}^\infty \mathbb{E}(\theta_{i,\gamma}^{(\theta_0)}) - \theta_* = \sum_{i=0}^\infty (I-\gamma\Sigma)^i(\theta_0-\theta_*) = (\gamma\Sigma)^{-1}(\theta_0-\theta_*) .$$

We can thus apply Theorem 3 to get a bound on $\mathbb{E}\left((\bar{\theta}_n^{(\gamma)} - \theta_*)(\bar{\theta}_n^{(\gamma)} - \theta_*)^\top\right)$. Indeed, with the previous notations, $\varphi = \bar{\theta}_n^{(\gamma)} - \theta_*$. We recall that:

$$\begin{aligned}\mathbb{E}(\Phi\Phi^\top) - (\mathbb{E}\Phi)(\mathbb{E}\Phi)^\top &= \frac{1}{n}\int_{\mathbb{R}^d}\left[\psi_\gamma(\theta)\psi_\gamma(\theta)^\top - (\psi_\gamma - \varphi)(\theta)(\psi_\gamma - \varphi)(\theta)^\top\right]\mathrm{d}\pi_\gamma(\theta) \\ &\quad + \frac{1}{n^2}\int_{\mathbb{R}^d}\left[\psi_\gamma(\theta)\psi_\gamma(\theta)^\top + \chi_\gamma^1(\theta) - \chi_\gamma^2(\theta)\right]\mathrm{d}\nu_0(\theta) + O(\rho^n)\,.\end{aligned}$$

Term proportional to $1/n$.
We need to compute the expectation under the stationary distribution of $\varphi(\theta)^{\otimes 2}$. For simplicity, we here denote $\mathbb{E}\varepsilon^{\otimes 2} = \int_{\mathbb{R}^d}\varepsilon_1(\theta)^{\otimes 2}\pi_\gamma(\mathrm{d}\theta)$. We have, according to Lemma 6:

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

The expectation of $\psi(\theta)\psi(\theta)^\top$ under the stationary is

$$\begin{aligned}\int_{\mathbb{R}^d}\psi(\theta)\psi(\theta)^\top\pi_\gamma(\mathrm{d}\theta) &= (\gamma\Sigma)^{-1}\gamma\left[\Sigma\otimes I + I\otimes\Sigma - \gamma\Sigma\otimes\Sigma\right]^{-1}\mathbb{E}\varepsilon^{\otimes 2}(\gamma\Sigma)^{-1} \\ &= \frac{1}{\gamma}(\Sigma^{-1}\otimes\Sigma^{-1})\left[\Sigma\otimes I + I\otimes\Sigma - \gamma\Sigma\otimes\Sigma\right]^{-1}\mathbb{E}\varepsilon^{\otimes 2}\,.\end{aligned}$$

Moreover,

$$\int_{\mathbb{R}^d}(\varphi(\theta) - \psi(\theta))(\varphi(\theta) - \psi(\theta))^\top\pi_\gamma(\mathrm{d}\theta) = \left[I - (\gamma\Sigma)^{-1}\right]\gamma\left[\Sigma\otimes I + I\otimes\Sigma\right]^{-1}\mathbb{E}\varepsilon^{\otimes 2}\left[I - (\gamma\Sigma)^{-1}\right].$$

Adding both these results and simplifying by $\left[\Sigma\otimes I + I\otimes\Sigma - \gamma\Sigma\otimes\Sigma\right]$, we get the following $1/n$-term:

$$\frac{1}{n}\mathbb{E}_{\theta\sim\pi_\gamma}\left[\psi(\theta)\psi(\theta)^\top - (R_\gamma\psi)(\theta)(R_\gamma\psi)(\theta)^\top\right] = \frac{1}{n}\Sigma^{-1}\Big[\int_{\mathbb{R}^d}\left(\varepsilon_1(\theta)^{\otimes 2}\right)\pi_\gamma(\mathrm{d}\theta)\Big]\Sigma^{-1}\,.$$

Term proportional to $1/n^2$.
We assume $\nu_0 = \delta_{\theta_0}$. This term is composed of three terms:

$$\begin{aligned}T_1 &:= -\mathbb{E}_{\theta_0\sim\nu_0}\left[\psi(\theta_0)\right]\mathbb{E}_{\theta_0\sim\nu_0}\left[\psi(\theta_0)\right]^\top \\ \psi(\theta_0) &= (\gamma\Sigma)^{-1}(\theta_0 - \theta_*) \\ T_1 &= -\frac{1}{\gamma^2}\Sigma^{-1}\left[(\theta_0 - \theta_*)^{\otimes 2}\right]\Sigma^{-1}\,.\end{aligned}$$

We note that, using $\psi = (\gamma\Sigma)^{-1}\varphi$, and $R_\gamma\psi = \psi - \varphi = -(I - (\gamma\Sigma)^{-1})\varphi$ that:

$$\begin{aligned}T_2 &:= \nu_0(\chi_\gamma^1) \\ &= (I - (\gamma\Sigma)^{-1})\nu_0(\chi_\gamma^3)(I - (\gamma\Sigma)^{-1})\,.\end{aligned}$$

Similarly:

$$\begin{aligned}T_2 &:= \nu_0(\chi_\gamma^1) \\ &= (\gamma\Sigma)^{-1}\nu_0(\chi_\gamma^3)(\gamma\Sigma)^{-1}\,.\end{aligned}$$

Where we recall that denote χ_γ^3 the solution to the Poisson equation associated with $\theta\mapsto\varphi(\theta)^{\otimes 2}$. We can compute explicitly this solution, indeed, following Equation 24:

$$\begin{aligned}\mathbb{E}\left[(\theta_{n,\gamma}^x - \theta_*)^{\otimes 2}\right] &= (I - \gamma\Sigma\otimes I - \gamma I\otimes\Sigma + \gamma^2 M)\mathbb{E}\left[(\theta_{n-1,\gamma}^x - \theta_*)^{\otimes 2}\right] + \mathbb{E}[\xi_n^{\otimes 2}] \\ \chi_\gamma^3(x) &:= \sum_{i=1}^\infty\mathbb{E}\left[(\theta_{n,\gamma}^x - \theta_*)^{\otimes 2}\right] - \pi_\gamma(\varphi(\theta)^{\otimes 2}) \\ &= (\gamma\Sigma\otimes I + \gamma I\otimes\Sigma - \gamma^2 M)^{-1}\left[\mathbb{E}\left[(\theta_{0,\gamma}^x - \theta_*)^{\otimes 2}\right] - \pi_\gamma(\varphi(\theta)^{\otimes 2}))\right] \\ \mathbb{E}_{\theta\sim\nu_0}\left[\chi_\gamma^3\right] &:= (\gamma\Sigma\otimes I + \gamma I\otimes\Sigma - \gamma^2 M)^{-1}\left[(\theta_0 - \theta_*)^{\otimes 2} - \pi_\gamma(\varphi(\theta)^{\otimes 2}))\right]\,.\end{aligned}$$

Simplification comes from the fact that we study an arithmetico-geometric recursion of the form $w_{n+1} = aw_n + b$, $a < 1$, and study $\sum_{i=0}^{\infty} w_n - w_\infty = (1-a)^{-1}(w_0 - w_\infty)$. Here we cannot apply the recursion with $(\Sigma \otimes I + I \otimes \Sigma - \gamma \Sigma \otimes \Sigma)$ because then b would depend on n. Finally,

$$\begin{aligned} T_2 + T_3 &= \frac{1}{\gamma}(\Sigma^{-1} \otimes \Sigma^{-1})(\Sigma \otimes I + I \otimes \Sigma - \gamma \Sigma \otimes \Sigma)\mathbb{E}_{\theta \sim \nu_0}\left[\chi(x)\right] \\ &= (\Sigma^{-1} \otimes \Sigma^{-1})\Omega\left[(\theta_0 - \theta_*)^{\otimes 2} - \gamma(\Sigma \otimes I + I \otimes \Sigma - \gamma M)^{-1}\mathbb{E}[\xi_1^{\otimes 2}])\right] . \end{aligned}$$

With: $\Omega = (\Sigma \otimes I + I \otimes \Sigma - \gamma \Sigma \otimes \Sigma)(\Sigma \otimes I + I \otimes \Sigma - \gamma T)^{-1}$.
Overall, we get that:

$$\begin{aligned} \mathbb{E}\bar{\theta}_n - \theta_* &= \frac{1}{n}(\gamma\Sigma)^{-1}(\theta_0 - \theta_*) \\ \mathrm{cov}(\bar{\theta}_n) &= \frac{1}{n}\Sigma^{-1}\big[\mathbb{E}\varepsilon^{\otimes 2}\big]\Sigma^{-1} - \frac{1}{n^2\gamma}\big[\Sigma^{-1} \otimes \Sigma^{-1}\big]\Omega\big[\Sigma \otimes I + I \otimes \Sigma - \gamma T\big]^{-1}\big[\mathbb{E}\xi^{\otimes 2}\big] \\ &\quad + \frac{1}{n^2}(\Sigma^{-1} \otimes \Sigma^{-1})(\Omega - I)(\theta_0 - \theta_*)^{\otimes 2} . \end{aligned}$$

Finally:

$$\begin{aligned} \mathbb{E}\left[(\bar{\theta}_n^{(\gamma)} - \theta_*)^{\otimes 2}\right] &= \frac{1}{n^2\gamma^2}(\Sigma^{-1} \otimes \Sigma^{-1})(\Omega)(\theta_0 - \theta_*)^{\otimes 2} + \frac{1}{n}\Sigma^{-1}\big[\mathbb{E}\varepsilon^{\otimes 2}\big]\Sigma^{-1} \\ &\quad - \frac{1}{n^2\gamma}\big[\Sigma^{-1} \otimes \Sigma^{-1}\big]\Omega\big[\Sigma \otimes I + I \otimes \Sigma - \gamma T\big]^{-1}\big[\mathbb{E}\xi^{\otimes 2}\big] . \end{aligned}$$

In the semi stochastic setting, we would get:

$$\begin{aligned} \mathbb{E}\left[(\bar{\theta}_n^{(\gamma)} - \theta_*)^{\otimes 2}\right] &= \frac{1}{n^2\gamma^2}(\Sigma^{-1} \otimes \Sigma^{-1})(\theta_0 - \theta_*)^{\otimes 2} + \frac{1}{n}\Sigma^{-1}\big[\mathbb{E}\varepsilon^{\otimes 2}\big]\Sigma^{-1} \\ &\quad - \frac{1}{n^2\gamma}\big[\Sigma^{-1} \otimes \Sigma^{-1}\big]\big[\Sigma \otimes I + I \otimes \Sigma - \gamma\Sigma \otimes \Sigma\big]^{-1}\big[\mathbb{E}\xi^{\otimes 2}\big] . \end{aligned}$$

□

C Further properties of the Markov chain $(\theta_k^{(\gamma)})_{k \geq 0}$

We give uniform bound on the moments of the chain $(\theta_k^{(\gamma)})_{k \geq 0}$ for $\gamma > 0$. We denote $\delta_n = \|\theta_n - \theta_*\|$. Denote by

$$\kappa = 2\mu L/(\mu + L) . \tag{33}$$

For $p \geq 1$ define

$$\mathrm{m}_p = \mathbb{E}^{1/p}\left[\|\varepsilon_1(\theta_*)\|^p\right] , \text{ for } p \geq 1 . \tag{34}$$

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

Lemma 11 (Final iterate)**.** *Under Assumptions **A**1,**A**2, **A**3, **A**7, one has the following bound on the $\mathbb{E}^{1/p}[\delta_{n+1}^{2p}]$, $p = 1, 2$. For the 2^{nd} order moment,*

$$\mathbb{E}[\delta_{n+1}^2] \leq (1 - 2\gamma\mu(1 - \gamma L))^n \delta_0^2 + \frac{\gamma\sigma^2}{\mu} . \tag{35}$$

For th 4^{th}-order moment, for $\gamma \leq \frac{1}{18L}$

$$\begin{aligned} \mathbb{E}^{1/2}[\delta_{n+1}^4] &\leq (1 - 2\gamma\mu(1 - 9\gamma L))\,\mathbb{E}^{1/2}[\delta_n^4] + 20\gamma^2\tau^2 \\ \mathbb{E}^{1/2}[\delta_n^4] &\leq (1 - 2\gamma\mu(1 - 9\gamma L))^n \mathbb{E}^{1/2}[\delta_0^4] + \frac{20\gamma\tau^2}{\mu} . \end{aligned}$$

*More generally, assume **A**1-**A**2-**A**3-**A**4(2p), for $p \geq 1$. There exist numerical constants C_p, D_p that only depend on p, such that, if $\gamma L \leq 1/2C_p$,*

$$\mathbb{E}_\theta^{1/p}\left[\left\|\theta_n^{(\gamma)} - \theta_*\right\|^{2p}\right] \leq (1 - 2\gamma\mu(1 - C_p\gamma L))^n \mathbb{E}_\theta^{1/p}\left[\|\theta_0 - \theta_*\|^{2p}\right] + \frac{D_p\gamma\mathrm{m}_{2p}^2}{\mu} .$$

Moreover, under stationary distribution π_γ, under the Assumptions above, one has:

$$\mathbb{E}_{\pi_\gamma}\left[\|\delta_n\|^{2p}\right] \leq \left(\frac{D_p \gamma \mathrm{m}_{2p}^2}{\mu}\right)^p . \tag{36}$$

Remark: Note that there is no contradiction between Equation (36) and Theorem 5, 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.

Lemma 11. We only prove the result for $p = 1, 2$ as it then naturally extends for any p.

The proof for the 2nd moment is very close to the one from (Needell et al., 2014) but we extend it without a.s. Lipschitzness (Assumption **A**4) but with Assumption **A**7. We recall that $\theta_{n+1} = \theta_n - \gamma f'(\theta_n) + \gamma \varepsilon_{n+1}$.

We have that

$$\|\theta_{n+1} - \theta_*\|^2 \quad = \quad \|\theta_n - \theta_* - \gamma f'(\theta_n) + \gamma \varepsilon_{n+1}\|^2 . \tag{37}$$

According to assumption **A**3, we have θ_n is $\mathcal{F}_n$ measurable, and $\mathbb{E}[\varepsilon_{n+1}|\mathcal{F}_n] = 0$. Thus $\mathbb{E}[\langle \theta_n - \theta_*, \varepsilon_{n+1}\rangle|\mathcal{F}_n] = 0$.

$$\begin{aligned}\mathbb{E}[\|\theta_{n+1} - \theta_*\|^2|\mathcal{F}_n] &= \mathbb{E}[\|\theta_n - \theta_*\|^2|\mathcal{F}_n] - 2\gamma\mathbb{E}[\langle f'(\theta_n), \theta_n - \theta_*\rangle|\mathcal{F}_n] \\ &\quad +\gamma^2\mathbb{E}[\|f_n'(\theta_n) - f_n'(\theta_*)\|^2|\mathcal{F}_n] + 2\gamma^2\mathbb{E}[\|f_n'(\theta_*)\|^2|\mathcal{F}_n] .\end{aligned} \tag{38}$$

Moreover, under Assumption **A**7, one has that $\mathbb{E}[\|f_n'(\theta_*)\|^2|\mathcal{F}_n] = \mathbb{E}[\|\varepsilon_1(\theta_*)\|^2] \leq \tau^2$ (using Hölder's inequality), and $\mathbb{E}[\|f_n'(\theta_n) - f_n'(\theta_*)\|^2|\mathcal{F}_n] \leq L\langle f'(\theta_n) - f'(\theta_*), \theta_n - \theta_*\rangle$. Thus:

$$\begin{aligned}\mathbb{E}[\delta_{n+1}^2|\mathcal{F}_n] &\leq \mathbb{E}[\delta_n^2|\mathcal{F}_n] - 2\gamma\langle f'(\theta_n) - f'(\theta_*), \theta_n - \theta_*\rangle + 2\gamma^2 L\langle f'(\theta_n) - f'(\theta_*), \theta_n - \theta_*\rangle \\ &\quad +\gamma^2\tau^2 \\ &\leq (1 - 2\gamma\mu(1 - \gamma L))\,\delta_n^2 + 2\gamma^2\tau^2 .\end{aligned} \tag{39}$$

Thus if $\gamma \leq \frac{1}{L}$, we have

$$\mathbb{E}[\delta_{n+1}^2] \quad \leq \quad (1 - 2\gamma\mu(1 - \gamma L))\,\mathbb{E}[\delta_n^2] + 2\gamma^2\tau^2 . \tag{40}$$

Thus if $\gamma L \leq 1$.

$$\mathbb{E}[\delta_{n+1}^2] \leq (1 - 2\gamma\mu(1 - \gamma L))^n\,\delta_0^2 + \gamma^2\tau^2\sum_{i=0}^{n-1}(1 - 2\gamma\mu)^i \tag{41}$$

$$= (1 - 2\gamma\mu(1 - \gamma L))^n\,\delta_0^2 + \frac{\gamma\tau^2}{\gamma\mu(1 - \gamma L)} . \tag{42}$$

□

Lemma 11. We have that

$$\begin{aligned}\delta_{n+1}^4 &= \left(\|\theta_n - \theta_*\|^2 - 2\gamma\langle f_n'(\theta_n), \theta_n - \theta_*\rangle + \gamma^2\|f_n'(\theta_n)\|^2\right)^2 \\ &= \left(\delta_n^2 - 2\gamma\langle f_n'(\theta_n), \theta_n - \theta_*\rangle + \gamma^2\|f_n'(\theta_n)\|^2\right)^2 \\ &= \delta_n^4 - 4\gamma\delta_n^2\langle f_n'(\theta_n), \theta_n - \theta_*\rangle + 4\gamma^2\langle f_n'(\theta_n), \theta_n - \theta_*\rangle^2 + 2\gamma^2\delta_n^2\|f_n'(\theta_n)\|^2 \\ &\quad -4\gamma^3\langle f_n'(\theta_n), \theta_n - \theta_*\rangle\|f_n'(\theta_n)\|^2 + \gamma^4\|f_n'(\theta_n)\|^4.\end{aligned}$$

Moreover:

$$\begin{aligned}\mathbb{E}[\|f_n'(\theta_n)\|^p|\mathcal{F}_n] &\leq 2^{p-1}\left(\mathbb{E}[\|f_n'(\theta_n) - f_n'(\theta_*)\|^p|\mathcal{F}_n] + \mathbb{E}[\|f_n'(\theta_*)\|^p|\mathcal{F}_n]\right) \\ &\leq 2^{p-1}\left(\|f_n'(\theta_n) - f_n'(\theta_*)\|^p + \mathbb{E}[\|\varepsilon_1(\theta_*)\|^p|\mathcal{F}_n]\right) \\ &\leq 2^{p-1}\left(\|f_n'(\theta_n) - f_n'(\theta_*)\|^p + \tau^p\right),\end{aligned} \tag{43}$$

using at the first line Minkowski's inequality and the fact that $x \mapsto x^p$ is convex on $\mathbb{R}^+$ for $p = 1, \ldots, 4$ thus $(x + y)^p \leq 2^{p-1}(x^p + y^p)$, and at the last line the Assumption **A**7 on the noise: $\mathbb{E}[\|\varepsilon_1(\theta_*)\|^p|\mathcal{F}_n] \leq \tau^p$.

Thus,

$$\begin{aligned}
\mathbb{E}[\delta_{n+1}^4|\mathcal{F}_n] &\leq \delta_n^4 - 4\gamma\delta_n^2\mathbb{E}[\langle f_n'(\theta_n), \theta_n - \theta_*\rangle|\mathcal{F}_n] + 4\gamma^2\mathbb{E}[\langle f_n'(\theta_n), \theta_n - \theta_*\rangle^2|\mathcal{F}_n] \\
&\quad +2\gamma^2\delta_n^2\mathbb{E}[\|f_n'(\theta_n)\|^2|\mathcal{F}_n] - 4\gamma^3\mathbb{E}[\langle f_n'(\theta_n), \theta_n - \theta_*\rangle\|f_n'(\theta_n)\|^2|\mathcal{F}_n] \\
&\quad +\gamma^4\mathbb{E}[\|f_n'(\theta_n)\|^4|\mathcal{F}_n] \\
&\leq \delta_n^4 - 4\gamma\delta_n^2\langle f'(\theta_n), \theta_n - \theta_*\rangle + 4\gamma^2\mathbb{E}[\|f_n'(\theta_n)\|^2\delta_n^2|\mathcal{F}_n] \\
&\quad +2\gamma^2\delta_n^2\mathbb{E}[\|f_n'(\theta_n)\|^2|\mathcal{F}_n] + 4\gamma^3\delta_n\mathbb{E}[\|f_n'(\theta_n)\|^3|\mathcal{F}_n] + \gamma^4\mathbb{E}[\|f_n'(\theta_n)\|^4|\mathcal{F}_n] \\
&\leq \delta_n^4 - 4\gamma\delta_n^2\langle f'(\theta_n), \theta_n - \theta_*\rangle + 12\gamma^2\delta_n^2\mathbb{E}[\|f_n'(\theta_n) - f_n'(\theta_*)\|^2|\mathcal{F}_n] \\
&\quad +16\gamma^3\delta_n\mathbb{E}[\|f_n'(\theta_n) - f_n'(\theta_*)\|^3|\mathcal{F}_n] + 8\gamma^4\mathbb{E}[\|f_n'(\theta_n) - f_n'(\theta_*)\|^4|\mathcal{F}_n] \\
&\quad +12\gamma^2\tau^2\delta_n^2 + 16\gamma^3\delta_n\tau^3 + 8\gamma^4\tau^4,
\end{aligned}$$

using Cauchy Schwartz several times for the second inequality and equation (43) for the third one.

Then, using part (ii) of Assumption **A**7:

$$\begin{aligned}
\mathbb{E}[\delta_{n+1}^4|\mathcal{F}_n] &\leq \delta_n^4 - 4\gamma\delta_n^2\langle f'(\theta_n), \theta_n - \theta_*\rangle + 12\gamma^2L\delta_n^2\langle f'(\theta_n), \theta_n - \theta_*\rangle \\
&\quad +16\gamma^3L^2\delta_n^2\langle f'(\theta_n), \theta_n - \theta_*\rangle + 8\gamma^4L^3\delta_n^2\langle f'(\theta_n), \theta_n - \theta_*\rangle \\
&\quad +12\gamma\tau^2\delta_n^2 + 8\gamma^2\tau^2\delta_n^2 + 8\gamma^4\tau^4 + 8\gamma^4\tau^4 \\
&= \delta_n^4 + (-4\gamma + 12\gamma^2L + 16\gamma^3L^2 + 8\gamma^4L^3)\delta_n^2\langle f'(\theta_n), \theta_n - \theta_*\rangle \\
&\quad +(12\gamma^2\tau^2 + 8\gamma^2\tau^2)\delta_n^2 + 16\gamma^4\tau^4 \\
&\leq \delta_n^4 - 4\gamma(1 - 9\gamma L)\delta_n^2\langle f'(\theta_n), \theta_n - \theta_*\rangle + 20\gamma^2\tau^2\delta_n^2 + 16\gamma^4\tau^4,
\end{aligned}$$

using $\gamma L \leq 1$ at the last line. Finally, using the smooth and strong convexity equation (15), we have:

$$\mathbb{E}[\delta_{n+1}^4|\mathcal{F}_n] \leq (1 - 4\gamma\mu(1 - 9\gamma L))\,\delta_n^4 + 20\gamma^2\tau^2\delta_n^2 + 16\gamma^4\tau^4,$$

Thus finally:

$$\begin{aligned}
\mathbb{E}[\delta_{n+1}^4] &\leq (1 - 4\gamma\mu(1 - 9\gamma L))\,\mathbb{E}[\delta_n^4] + 20\gamma^2\tau^2\mathbb{E}[\delta_n^2] + 16\gamma^4\tau^4 \\
&\leq \left((1 - 4\gamma\mu(1 - 9\gamma L))^{1/2}\,\mathbb{E}[\delta_n^4]^{1/2} + 20\gamma^2\tau^2\right)^2.
\end{aligned}$$

Using that $20\gamma^2\tau^2\mathbb{E}[\delta_n^2] \leq (1 - 4\gamma\mu(1 - 9\gamma L))^{1/2}\mathbb{E}[\delta_n^4]^{1/2}40\gamma^2\tau^2$ *i.e.*, $\mathbb{E}[\delta_n^2] \leq \mathbb{E}[\delta_n^4]^{1/2}$, and $(1 - 4\gamma\mu(1 - 9\gamma L))^{1/2} \geq 1/2$ which is true if $\gamma \leq \frac{1}{9L}$ and $(1 - 4\gamma\mu(1 - 9\gamma L)) \geq (1 - 4/9)^{1/2} \geq 1/2$.

$$\mathbb{E}^{1/2}[\delta_{n+1}^4] \leq (1 - 2\gamma\mu(1 - 9\gamma L))\,\mathbb{E}^{1/2}[\delta_n^4] + 20\gamma^2\tau^2.$$

If $9\gamma L \leq 1$.

Which concludes the proof. □

Theorem 12. *Assume* ***A****1-****A****2-****A****3-****A****4(2k₂)-****A****9(k₁)- for $k_1, k_2 \in \mathbb{N}$, $k_1 \geq 1$. Let $g : \mathbb{R}^d \to \mathbb{R}$ satisfying* ***A****5(k₁, k₂) for $k_2 \in \mathbb{N}$. Then, there exists $C_{k_2} \geq 0$ only depending on k_2 such that for all $\gamma \in (0, C_{k_2}/L)$, for all initial point $\theta \in \mathbb{R}^d$, there exists C such that for all $n \geq 1$:*

$$\left|\mathbb{E}_\theta\left[n^{-1}\sum_{i=1}^n \left\{g(\theta_i^{(\gamma)})\right\}\right] - \int_{\mathbb{R}^d} g(\theta)\pi_\gamma(\mathrm{d}\theta)\right| \leq Cn^{-1}\,.$$

Proof.

$$\begin{aligned}
\left|\sum_{i=1}^n \left(\mathbb{E}_\theta\left[g(\theta_{i,\gamma}^\theta)\right] - \int_{\mathbb{R}^d} g(\theta)\pi_\gamma(\mathrm{d}\theta)\right)\right| &= \sum_{i=1}^n \left|\left(\int_{y\in\mathbb{R}^d} \mathbb{E}_\theta\left[g(\theta_{i,\gamma}^\theta) - g(\theta_{i,\gamma}^y)\right]\pi_\gamma(y)\right)\right| \\
&= \sum_{i=1}^n \left(\int_{y\in\mathbb{R}^d} \mathbb{E}_\theta\left[\left\|g(\theta_{i,\gamma}^\theta) - g(\theta_{i,\gamma}^y)\right\|\right]\pi_\gamma(y)\right).
\end{aligned}$$

Using Lemma 14, a.s.,

$$\left\|g(\theta_{i,\gamma}^\theta) - g(\theta_{i,\gamma}^y)\right\| \leq a_g\left\|\theta_{i,\gamma}^\theta - \theta_{i,\gamma}^y\right\|\left(\left(b_g + \left\|\theta_{i,\gamma}^\theta - \theta_*\right\|^{k_2} + \left\|\theta_{i,\gamma}^y - \theta_*\right\|^{k_2}\right)\right).$$

By Cauchy Schwartz, then Minkowski:

$$\begin{aligned}\mathbb{E}_\theta\left[\left\|g(\theta^\theta_{i,\gamma})-g(\theta^y_{i,\gamma})\right\|\right] \leq & a_g \mathbb{E}_\theta^{1/2}\left[\left\|\theta^\theta_{i,\gamma}-\theta^y_{i,\gamma}\right\|^2\right]\mathbb{E}_\theta^{1/2}\left[(b_g+\left\|\theta^\theta_{i,\gamma}-\theta_*\right\|^{k_2}+\left\|\theta^y_{i,\gamma}-\theta_*\right\|^{k_2})^2\right] \\ \leq & a_g\left(W_2(R^n_\gamma(\theta,.),R^n_\gamma(y,.)\right)^{1/2} \\ & \times\left(b_g+\mathbb{E}_\theta^{1/2}\left[\left\|\theta^\theta_{i,\gamma}-\theta_*\right\|^{2k_2}\right]+\mathbb{E}_\theta^{1/2}\left[\left\|\theta^y_{i,\gamma}-\theta_*\right\|^{2k_2}\right]\right) .\end{aligned}$$

With $\rho=(1-\gamma\mu(1-\gamma L))$, we have, using Lemma 11 , which implies that:

$$\mathbb{E}_\theta^{1/2}\left[\left\|\theta^{(\gamma)}_n-\theta_*\right\|^{2p}\right]\leq 2^{p/2-1}\mathbb{E}_\theta^{1/2}\left[\left\|\theta^{(\gamma)}_0-\theta_*\right\|^{2p}\right]+2^{p/2}\left(\frac{D_p\gamma \mathrm{m}^2_{2p}}{\mu}\right)^{p/2} .$$

$$\begin{aligned}\mathbb{E}_\theta\left[\left\|g(\theta^\theta_{i,\gamma})-g(\theta^y_{i,\gamma})\right\|\right] \leq & a_g\rho^{n/2}\left\|\theta-y\right\|\left(b_g+2^{p/2-1}\mathbb{E}_\theta^{1/2}\left[\left\|\theta^{(\gamma)}_0-\theta_*\right\|^{2k_2}\right]\right. \\ & \left.+2^{p/2-1}\left\|y-\theta_*\right\|^{k_2}2^{p/2+1}\left(\frac{D_p\gamma \mathrm{m}^2_{2p}}{\mu}\right)^{p/2}\right) .\end{aligned}$$

Thus

$$\left|\mathbb{E}_\theta\left[n^{-1}\sum_{i=1}^n\left\{g(\theta^{(\gamma)}_i)\right\}\right]-\int_{\mathbb{R}^d}g(\theta)\pi_\gamma(\mathrm{d}\theta)\right|\leq\frac{C}{n}\sum_{i=1}^n\rho^{n/2}\leq\frac{C}{\gamma\mu n}$$

$$\begin{aligned}C = & a_g\int_{\mathbb{R}^d}\left(\left\|\theta-y\right\|\left(b_g+2^{p/2-1}\mathbb{E}_\theta^{1/2}\left[\left\|\theta^{(\gamma)}_0-\theta_*\right\|^{2k_2}\right]+2^{p/2-1}\left\|y-\theta_*\right\|^{k_2}\right.\right. \\ & \left.\left.2^{p/2+1}\left(\frac{D_p\gamma \mathrm{m}^2_{2p}}{\mu}\right)^{p/2}\right)\mathrm{d}\pi_\gamma(y)\right) .\end{aligned}$$

□

D Regularity of the gradient flow and estimates on Poisson solution

Let $k\in\mathbb{N}^*$ and consider the following assumption.

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

Lemma 13. *Assume **A**1 and **A**9$(k+1)$ for $k\in\mathbb{N}$, $k\geq 1$.*

a) For all $t\geq 0$, $\phi_t\in C^k(\mathbb{R}^d)$. In addition for all $\theta\in\mathbb{R}$, $\phi^{(k)}_t(x):t\mapsto D^k\phi_t(\theta)$ satisfies the following ordinary differential equation,

$$\dot{\phi}^{(k)}_t(x)=D^k\left\{\nabla f(\phi_t(\theta))\right\}\ ,\ \text{for all } t\geq 0\ ,$$

with $\phi^{(2)}_0(x)=\mathrm{Id}$ and $\phi^{(k)}_0(x)=0$ for $k\geq 2$.

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

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

$$\nabla\phi_t(\theta_*)=\mathrm{e}^{-\nabla^2 f(\theta_*)} .$$

d) If $k\geq 3$, for all $t\geq 0$ and $i,j,k\in\{1,\ldots,d\}$,

$$\left\langle D^2\phi_t(\theta_*)\left\{\mathbf{v}_i,\mathbf{v}_j\right\},\mathbf{v}_k\right\rangle=\frac{\mathrm{e}^{-\lambda_i t}-\mathrm{e}^{-(\lambda_k+\lambda_j)t}}{\lambda_i-\lambda_k-\lambda_j} ,$$

where $\{\mathbf{v}_1,\ldots,\mathbf{v}_d\}$ and $\{\lambda_1,\ldots,\lambda_d\}$ are the eigenvectors and the eigenvalues of $\nabla^2 f(\theta_*)$ respectively satisfying for all $i\in\{1,\ldots,d\}$, $\nabla^2 f(\theta_*)\mathbf{v}_i=\lambda_i\mathbf{v}_i$.

Proof. a) This is a fundamental result on the regularity of flows of autonomous differential equations, see e.g. (Hartman, 1982, Theorem 4.1 Chapter V)

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

c) By Lemma 13-a) and since θ_* is an equilibrium point, for all $x \in \mathbb{R}^d$, $\xi_t^x(\theta_*) = D\phi_t(\theta_*)\{x\}$ satisfies the following ordinary differential equation

$$\dot{\xi_s^x}(\theta_*) = -\nabla^2 f(\phi_s(\theta_*))\xi_s^x(\theta_*)\mathrm{d}s = -\nabla^2 f(\theta_*)\xi_s^x(\theta_*)\mathrm{d}s \; . \tag{44}$$

with $\xi_0^x(\theta_*) = x$. The proof then follows from uniqueness of the solution of (44).

d) By Lemma 13-a), for all $x_1, x_2 \in \mathbb{R}^d$, $\xi_t^{x_1,x_2}(\theta_*) = D^i\phi_t(\theta_*)\{x_1 \otimes x_2\}$ satisfies the ordinary stochastic differential equation:

$$\frac{\mathrm{d}\xi_s^{x_1,x_2}}{\mathrm{d}s}(\theta_*) = -D^3 f(\phi_s(\theta_*))\{\nabla\phi_s(\theta_*)x_1 \otimes \nabla\phi_s(\theta_*)x_2 \otimes \mathbf{e}_i\} - D^2 f(\phi_s(\theta_*))\{\xi_s^{x_1,x_2}\}\mathbf{e}_i \; .$$

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

$$\frac{\mathrm{d}\xi_s^{x_1,x_2}}{\mathrm{d}s}(\theta_*) = -D^3 f(\theta_*)\left\{\mathrm{e}^{-\nabla^2 f(\theta_*)t}x_1 \otimes \mathrm{e}^{-\nabla^2 f(\theta_*)t}x_2 \otimes \mathbf{e}_i\right\} - D^2 f(\theta_*)\{\xi_s^{x_1,x_2}\}\mathbf{e}_i \; .$$

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

$$\frac{\mathrm{d}\left\langle \xi_s^{\mathbf{v}_i,\mathbf{v}_j}, \mathbf{v}_k \right\rangle}{\mathrm{d}s} = -D^3 f(\theta_*)\left\{\mathrm{e}^{-\lambda_i t}\mathbf{v}_i \otimes \mathrm{e}^{-\lambda_j t}\mathbf{v}_j \otimes \mathbf{v}_k\right\} - \lambda_k \left\langle \xi_s^{\mathbf{v}_i,\mathbf{v}_j}, \mathbf{v}_k \right\rangle \; .$$

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

□

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

$$h_g(\theta) = \int_0^\infty (g(\phi_s(\theta)) - g(\theta_*))dt \; . \tag{45}$$

Note that h_g is well-defined by Lemma 13-b) and since g is assumed to be locally-Lipschitz. Note that by (8), we have for all $g : \mathbb{R}^d \to \mathbb{R}$, locally Lipschitz,

$$\mathcal{A}h_g(\theta) = g(\theta) - g(\theta_*) \; . \tag{46}$$

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

$$h_{\mathrm{Id}}(\theta) = \int_0^\infty \{\phi_s(\theta) - \theta_*\}\, dt \; . \tag{47}$$

Note that h_{Id} is also well-defined by Lemma 13-b).

Lemma 14. *Let $g : \mathbb{R}^d \to \mathbb{R}$ satisfying $\boldsymbol{A}5(k_1, k_2)$ for $k_1, k_2 \in \mathbb{N}$, $k_1 \geq 1$.*

a) Then for all $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$|g(\theta_1) - g(\theta_2)| \leq a_g \|\theta_1 - \theta_2\| \left\{b_g + \|\theta_1 - \theta_*\|^{k_2} + \|\theta_2 - \theta_*\|^{k_2}\right\} \; .$$

Assume in addition $\boldsymbol{A}1$ and $\boldsymbol{A}9(k_1 + 1)$.

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

$$|h_g|(\theta) \leq a_g \left\{(b_g/\mu)\|\theta - \theta_*\| + (k_2\mu)^{-1}\|\theta - \theta_*\|^{k_2}\right\} \; .$$

c) If $k_1 \geq 2$, then $\nabla h_{\mathrm{Id}}(\theta_*) = (\nabla^2 f(\theta_*))^{-1}$. If $k_1 \geq 3$, then for all $i, j \in \{1, \ldots, d\}$,

$$\frac{\partial^2 h_{\mathrm{Id}}}{\partial\theta_i \partial\theta_j}(\theta_*) = -D^3 f(\theta_*) \left\{ \left[\left(\nabla^2 f(\theta_*) \otimes \mathrm{Id} + \mathrm{Id} \otimes \nabla^2 f(\theta_*) \right)^{-1} \{\mathbf{e}_i \otimes \mathbf{e}_j\} \right] \otimes \mathbf{e}_i \right\} (\nabla^2 f(\theta_*))^{-1} \mathbf{e}_i \ ,$$

where $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$ are the canonical basis of $\mathbb{R}^d$.

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

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

The proof is then concluded using $\mathbf{A}5(k_1, k_2)$ and

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

b) For all $\theta \in \mathbb{R}^d$, we have using the first result of the Lemma and (45)

$$|h_g(\theta)| \leq a_g \int_0^{+\infty} \|\phi_s(\theta) - \theta_*\| \left\{ b_g + \|\phi_s(\theta) - \theta_*\|^{k_2} \right\} \mathrm{d}s \ .$$

The proof then follows from Lemma 13-b).

c) The proof is a direct consequence of Lemma 13-c)-d) and (45).

□

Theorem 15. *Assume **A**1-**A**9(k_1+1) for $k_1, k_2 \in \mathbb{N}$, $k_1 \geq 2$. Let $g : \mathbb{R}^d \to \mathbb{R}$ satisfying **A**5(k_1, k_2) for $k_2 \in \mathbb{N}$.*

a) For all $t \geq 0$, $\phi_t \in C^{k_1}(\mathbb{R}^d)$ and for all $i \in \{1, \ldots, k\}$, there exists $C_i \geq 0$ such that for all $\theta \in \mathbb{R}^d$ and $t \geq 0$,

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

b) Let $g \in C^{k_1}(\mathbb{R}^d)$. Then $h_g \in C^{k_1}(\mathbb{R}^d)$ and for all $i \in \{0, \ldots, k_1\}$, there exists $C_i \geq 0$ such that for all $\theta \in \mathbb{R}^d$,

$$\left\| D^i h_g(\theta) \right\| \leq C_i \left\{ 1 + \|\theta - \theta_*\|^{k_2} \right\} \ .$$

Proof. a) The proof is by induction on k_1. By Lemma 13-a), for all $x \in \mathbb{R}^d$, and $\theta \in \mathbb{R}^d$, $\xi_t^x(\theta) = D\phi_t(\theta)\{x\}$ satisfies

$$\frac{\mathrm{d}\xi_s^x}{\mathrm{d}s}(\theta) = -\nabla^2 f(\phi_s(\theta)) \xi_s^x(\theta) \mathrm{d}s \ . \tag{48}$$

with $\xi_0^x(\theta) = x$. Now differentiating $s \to \|\xi_s^x(\theta)\|^2$, using $\mathbf{A}1$ and Grünwall's inequality, we get $\|\xi_s^x(\theta)\|^2 \leq \mathrm{e}^{-2ms} \|x\|^2$ which implies the result for $k_1 = 2$.

Let now $k_1 > 2$. Using again Lemma 13-a), Faà di Bruno's formula (Levy, 2006, Theorem 1) and since (7) can be written on the form

$$\frac{\mathrm{d}\phi_t}{\mathrm{d}s}(\theta) = -\sum_{j=1}^{d} Df(\phi_t(\theta)) \{e_i\} e_i \ ,$$

for all $i \in \{2, \ldots, k_1\}$, $\theta \in \mathbb{R}^d$ and $x_1, \cdots, x_i \in \mathbb{R}^d$, $\xi_t^{x_1, \cdots, x_i}(\theta) = D^i \phi_t(\theta) \{x_1 \otimes \cdots \otimes x_i\}$ satisfies the ordinary differential equation:

$$\frac{\mathrm{d}\xi_s^{x_1, \cdots, x_i}}{\mathrm{d}s}(\theta) = -\sum_{j=1}^{d} \sum_{\Omega \in \mathsf{P}(\{1, \ldots, i\})} D^{|\Omega|+1} f(\phi_s(\theta)) \left\{ e_i \otimes \bigotimes_{l=1}^{i} \bigotimes_{j_1, \cdots, j_l \in \Omega} \xi_s^{x_{j_1}, \cdots, x_{j_l}}(\theta) \right\} e_i \ , \tag{49}$$

where $\mathsf{P}(\{1, \ldots, i\})$ is the set of partitions of $\{1, \ldots, i\}$, which does not contain the empty set and $|\Omega|$ is the cardinal of $\Omega \in \mathsf{P}(\{1, \ldots, i+1\})$. We now show by induction on i that for all $i \in \{1, \ldots, k_1\}$, there exists a universal constant C_i such that for all $t \geq 0$ and $\theta \in \mathbb{R}^d$,

$$\sup_{x \in \mathbb{R}^d} \left\| D^i \phi_t(\theta) \right\| \leq C_i \mathrm{e}^{-\mu t} \ . \tag{50}$$

For $i=1$, the result follows from the case $k_1=1$. Assume that the result is true for $\{1,\dots,i\}$ for $i\in\{1,\dots,k_1-1\}$. We show the result for $i+1$. By (49), we have for all $\theta\in\mathbb{R}^d$ and $x_1,\cdots,x_i\in\mathbb{R}^d$,

$$\frac{\left\|\xi_t^{x_1,\cdots,x_{i+1}}(\theta)\right\|^2}{\mathrm{d}t} = -\int_0^t \sum_{\Omega\in\mathsf{P}(\{1,\dots,i+1\})} D^{|\Omega|+1}f(\phi_s(\theta))\left\{\xi_t^{x_1,\cdots,x_{i+1}}(\theta)\otimes\bigotimes_{l=1}^{i+1}\bigotimes_{j_1,\dots,j_l\in\Omega}\xi_s^{x_{j_1},\cdots,x_{j_l}}(\theta)\right\}\mathrm{d}s\ .$$

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 $x\in\mathbb{R}^d$ (50) holds for $i+1$.

b) The proof is a consequence of a), (45), $\mathbf{A}5(k_1,k_2)$ and Leibniz's rule.

□

E Proof of Theorem 5

We preface the proof of the Theorem by two fundamental first estimates.

Theorem 16. *Assume **A**1-**A**2-**A**3-**A**4(2(k₂+3)), for $k_1,k_2\in\mathbb{N}$, $k_1\geq 1$. Let $g:\mathbb{R}^d\to\mathbb{R}$ satisfying **A**5(3,k₂). Then, there exists $C_{k_2}\geq 0$ only depending on k_2 such that for all $\gamma\in(0,C_{k_2}/L)$, $n\in\mathbb{N}^*$, $\gamma>0$ and $\theta\in\mathbb{R}^d$,*

$$\mathbb{E}_\theta\left[n^{-1}\sum_{i=1}^n\left\{g(\theta_i^{(\gamma)})-g(\theta_*)\right\}\right] = \frac{\mathbb{E}_\theta\left[h_g(\theta_{n+1}^{(\gamma)})\right]-h_g(\theta)}{n\gamma} - (\gamma/2)\int_{\mathbb{R}^d}D^2h_g(\tilde\theta)\mathbb{E}\left[\left\{\varepsilon(\tilde\theta)\right\}^{\otimes 2}\right]\mathrm{d}\pi_\gamma(\tilde\theta)+(\gamma/n)\tilde A_1(\theta)+\gamma^2\tilde A_2(\theta,n)\ ,$$

where

$$\tilde A_1(\theta)\leq C\left\{1+\|\theta-\theta_*\|^{k_2+2}\right\}\ ,\tilde A_2(\theta,n)\leq C\left\{1+\|\theta-\theta_*\|^{k_2+3}/n\right\}\ ,$$

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

Proof. Let $n\in\mathbb{N}^*$, $\gamma>0$ 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 15 shows that $h_g\in C^3(\mathbb{R}^d)$. Therefore using (1) and the Taylor expansion formula, we have for all $i\in\{1,\dots,n\}$

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

where $s_i^{(\gamma)}\in[0,1]$ and $\Delta\theta_{i+1}^{(\gamma)}=\theta_{i+1}^{(\gamma)}-\theta_i^{(\gamma)}$. Therefore by (46), we get

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

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

$$\mathbb{E}_\theta\left[n^{-1}\sum_{i=1}^{n}\left\{g(\theta_i^{(\gamma)})-g(\theta_*)\right\}\right]=\frac{\mathbb{E}_\theta\left[h_g(\theta_{n+1}^{(\gamma)})\right]-h_g(\theta)}{n\gamma} - (\gamma/2)\int_{\mathbb{R}^d} D^2h_g(\tilde\theta)\mathbb{E}\left[\left\{\varepsilon(\tilde\theta)\right\}^{\otimes 2}\right]\mathrm{d}\pi_\gamma(\tilde\theta)+\tilde A_1+\tilde A_2\ ,$$

where

$$\tilde A_1=(\gamma/(2n))\mathbb{E}_\theta\left[\sum_{i=1}^{n}\left(D^2h_g(\theta_*)\left\{\varepsilon_{i+1}(\theta_*)\right\}^{\otimes 2}-D^2h_g(\theta_i^{(\gamma)})\left\{-\nabla f(\theta_i^{(\gamma)})+\varepsilon_{i+1}(\theta_i^{(\gamma)})\right\}^{\otimes 2}\right)\right]$$

$$\tilde A_2=-(\gamma^2/(3!n))\mathbb{E}_\theta\left[\sum_{i=1}^{n}D^3h_g(\theta_i^{(\gamma)}+s_i^{(\gamma)}\Delta\theta_{i+1}^{(\gamma)})\left\{-\nabla f(\theta_i^{(\gamma)})+\varepsilon_{i+1}(\theta_i^{(\gamma)})\right\}^{\otimes 3}\right]\ .$$

The proof is then concluded using Theorem 15, Lemma 11 and Theorem 12. □

Corollary 17. *Assume **A**1-**A**2-**A**3-**A**4(2(k₂+3)), for $k_1,k_2\in\mathbb{N}$, $k_1\geq 1$. Let $g:\mathbb{R}^d\to\mathbb{R}$ satisfying **A**5(3,k₂). Then there exists $C_{k_2}\geq 0$ only depending on k_2 such that for all $\gamma\in(0,C_{k_2}/L)$, there exists $C\geq 0$ independent of γ such that*

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

Proof. The proof is a direct consequence of Theorem 12 and Theorem 16. □

Proof of Theorem 5. Under the stated assumptions, $\theta\mapsto D^2h_g(\theta)\mathbb{E}\left[\left\{\varepsilon(\theta)\right\}^{\otimes 2}\right]$ satisfies the conditions of Corollary 17. The proof then follows from combining Corollary 17 applied to this function and Theorem 16. □

References

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

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

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

D. Bertsekas. *Nonlinear programming*. Athena Scientific, 1995.

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

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

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

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

A. Dieuleveut, N. Flammarion, and F. Bach. Harder, Better, Faster, Stronger Convergence Rates for Least-Squares Regression. *ArXiv e-prints*, Feb. 2016.

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

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

P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Parallelizing Stochastic Approximation Through Mini-Batching and Tail-Averaging. *ArXiv e-prints*, Oct. 2016.

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

G. L. Jones. On the Markov chain central limit theorem. *Probability Surveys*, 1:299–320, 2004.

H. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.

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

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

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

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

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

S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer-Verlag Inc, Berlin; New York, 1993.

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

D. Needell, R. Ward, and N. Srebro. Stochastic Gradient Descent, Weighted Sampling, and the Randomized Kaczmarz algorithm. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1017–1025. Curran Associates, Inc., 2014.

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

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer, 2004.

Y. Nesterov and J. P. Vial. Confidence Level Solutions for Stochastic Programming. *Automatica*, 44(6):1559–1568, June 2008.

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

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

A. Rakhlin, O. Shamir, and K. Sridharan. Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. *ArXiv e-prints*, Sept. 2011.

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

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

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

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

O. Shamir and T. Zhang. Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes. *Proceedings of the 30^{t}h International Conference on Machine Learning*, 2013.

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

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

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

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

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