A LECTURE ON SGD AND MARKOV CHAINS
based on a paper by Dieuleveut, Durmus, and Bach (2020)

SIMONE MARIA GIANCOLA*†

December 17, 2024

CONTENTS

1 INTRODUCTION

In this short lecture, we will review the paper in the title. After presenting the framework, we will prove the starting statement, and just briefly comment the other ones, which have very long proofs. To conclude, we will overview the Richardson-Romberg extrapolation method, which is one direct application of the results, and the ideas behind the deeper theorems.
Emphasis is on intuition and quick understanding. References are to a minimum. Computations, when performed, are explicit. Throughout, we omit the full expressions of the theorems, but specify when these are available in the original publication.

NOTATION We use bold for vectors, and curly latex for random variables. For example $\boldsymbol{\theta}$ is a random vector, while $\boldsymbol{\theta}^\star$ will be our deterministic optimum. The rest of the symbols are either standard or defined when discussed first. The main takeaway is this explicit distinction between what is random and what is not.

1.I *Motivation and Setting*

Stochastic gradient descent (SGD) is a standard tool in machine learning (ML). However, the choice of the step-size is classical only in the deterministic case, where the algorithm degrades to gradient descent (GD). If we train over random samples, there is a *gap* between theory and practice:

*Université Paris-Saclay, Orsay
†email: simonegiancola09@gmail.com

- a scaling $O(1/k)$ is advisable, while still being non-robust to ill-conditioning;
- experiments suggest $O(1/\sqrt{k})$ and averaging.

We wish to tackle this issue by providing intuitions on a set of results that indeed *bridges the gap*. In the spirit of isolating phenomena, let us choose constant step-size. The immediate justifications are that: (i) it is easier theoretically; (ii) there are fewer parameters to optimize (one less); (iii) initial conditions are anyway forgotten exponentially fast for the problems we consider; (iv) in practice, it gets close to the global optimum, and in ML we do not care when we are already at machine precision.
Let us introduce the notation. In this report SGD is expressed via the recursion:

$$\theta_{k+1}^{(\gamma)} = \theta_k^{(\gamma)} - \gamma \left[\nabla f(\theta_k^{(\gamma)}) + \epsilon_{k+1}(\theta_k^{(\gamma)}) \right], \qquad \theta_0 \sim \lambda, \tag{1.1}$$

(learning rate in $\mathbb{R}_+$; true gradient; new estimator; noise)

where λ is an initial distribution and $f : \mathbb{R}^d \to \mathbb{R}$ is the function we wish to minimize (a loss function). It will be nice and admit a global minimum $\theta^\star$.[1]

Main Observation

The sequence $(\theta_k^{(\gamma)})_{k\in\mathbb{N}}$ is a homogeneous Markov chain.

To remedy these oscillations, we consider a running mean:

$$\overline{\theta}_k^{(\gamma)} = \frac{1}{k+1} \sum_{j=0}^{k} \theta_j^{(\gamma)}. \tag{1.2}$$

We will prove under appropriate conditions on the sequence of iterates that a CLT holds, so that the running mean converges to the mean of the stationary distribution at a rate $O(1/k)$. The mean of the stationary distribution is obviously:

$$\overline{\theta}_\gamma = \int_{\mathbb{R}^d} \xi \mathrm{d}\pi_\gamma(\xi). \tag{1.3}$$

Therefore, we can split the deviation of this running mean from the real (global) optimum of the function into:

$$\left\| \overline{\theta}_k^{(\gamma)} - \theta^\star \right\|_2^2 \leqslant \left\| \overline{\theta}_k^{(\gamma)} - \overline{\theta}_\gamma \right\|_2^2 + \left\| \overline{\theta}_\gamma - \theta^\star \right\|_2^2. \tag{1.4}$$

(stochastic error; deterministic error)

1.II *Summary of results*

We will briefly argue that for quadratic functions the second term is null, but in general the oscillations are of order $O(\gamma)$. Then, we will derive an **explicit asymptotic expansion** of these oscillations, i.e. an expansion of $\overline{\theta}_\gamma - \theta^\star$ in the parameters of the algorithm. Like in classic ML, through a bias-variance decomposition, a quantitative CLT to expand the stochastic part $\mathbb{E}\left[\left\| \theta_k^{(\gamma)} - \overline{\theta}_\gamma \right\|_2^2 \right]$ is derived. In particular the bias depends on the initial conditions and the variance depends on the structure of the noise. Building further on that, we will discuss a **non-asymptotic** expansion between the stationary distribution π_γ and the Dirac at $\theta^\star$ in terms of the step-size. This means in practice that for a nice class of functions g, the integral with respect to the stationary measure is decomposed in Taylor-style centered at $g(\theta^\star)$. All together, these results will describe the invariant distribution and how it is reached from any measure. In doing so, we will also justify a nice numerical trick,

[1] The original paper is more general and can work with tensors, but we avoid this for the sake of simplicity.

the Richardson-Romberg extrapolation, and find the excuse to discuss how to recover concentration from implicit functions depending on the problem.

Advantages

(A1) Asymptotic expansions are explicit in the parameters! Full quantitative phenomenology.

(A2) We can build confidence intervals for $\boldsymbol{\theta}^\star$ (Chen et al. 2020; Su and Zhu 2018).

(A3) More informed design of automatic restart schemes.

1.III *Related work*

We keep references here to a minimum, and refer to the original publication for valuable comments.
The most direct relative of this work is (Aguech, Moulines, and Priouret 2000), which develops a theory that matches the results for linear regression. For the perspective of SGD as a discretized gradient flow and its connections to Markov chains theory there are many important works. In particular, Fort and Pagès (1999) find tightness of the invariant distribution in a neighborhood of γ small and invariance with respect to the gradient flow of the limit distribution. However, they assume the process is Feller, while this work rather puts assumptions on the objective function (namely, strong convexity).
For convergence of SGD, there are many works studying the bias variance trade-off, with an emphasis on the dependence on initial conditions. Concerning the perspective of comparing the discretized invariant distribution and the continuous version and the Richardson-Romberg trick there are earlier works that combined the two.
The first part of corollary 2.10 matches the result of Defossez and Bach 2015, while proposition 2.6 is in part an extension of the work of Ljung, Pflug, and Walk (2012), in the sense that it quantifies their result that $\sqrt{\gamma}(\pi_\gamma - \delta_{\boldsymbol{\theta}^\star})_{\gamma>0} \overset{\gamma\to 0}{\underset{\mathrm{d}}{\to}} \mathcal{N}(0,1)$.

2 RESULTS

2.I *Assumptions*

We present here a quick overview of the assumptions, and postpone explicit writing to the appendix A.

(on f) μ-strong convexity, in $\mathrm{C}^5(\mathbb{R}^d,\mathbb{R})$ with uniform bounds,[2] with L-co-coercive derivative (see appendix D for a refresher).

(on $(\boldsymbol{\epsilon}_k)_{k\in\mathbb{N}}$) The sequence of noise terms is adapted wrt to a filtration and is a Markov chain wrt to it, the p^{th} norm of the noise is controlled by τ_p and the covariance of the noise is $\mathrm{C}^3(\mathbb{R}^d,\mathbb{R})$ with operator norms bounded by $M_\epsilon(1+\|\boldsymbol{\theta}-\boldsymbol{\theta}^\star\|_2^{k_\epsilon})$ when $\boldsymbol{\epsilon}\equiv\boldsymbol{\epsilon}(\boldsymbol{\theta})$.

(misc) The initial iterate is measurable $\theta_0\in\mathcal{F}_0$, and we have access to unbiased estimates, but the noise may depend on the current iterate, i.e. for all $\boldsymbol{\theta}\in\mathbb{R}^d$:

$$\nabla \mathrm{f}_{k+1}(\boldsymbol{\theta}) = \nabla f(\boldsymbol{\theta}) + \boldsymbol{\epsilon}_{k+1}, \qquad , \boldsymbol{\epsilon}_{k+1}\equiv\boldsymbol{\epsilon}_{k+1}(\boldsymbol{\theta}). \tag{2.1}$$

Here $\nabla \mathrm{f}$ is effectively the random gradient of f which is random and ∇f is the true gradient at $\boldsymbol{\theta}$. In particular, this implies that the noise may be not i.i.d. and this makes $\theta_k\in\mathcal{F}_k$.

2.II *A baby example*

Let us degrade the generality of the paper to a practical scenario, which is a special case discussed in (Dieuleveut, Durmus, and Bach 2020, example 1). Consider a loss $\mathcal{L}:\mathbb{R}^d\times\mathbb{R}\times\mathbb{R}^d\to\mathbb{R}$ that maps the triplet of feature vector, observed signal and proposed latent vector $(x,y,\boldsymbol{\theta})$ into a scalar. If we allow the features and the labels to be random, as when they come from a dataset $\mathscr{D}$ of observations, then we are interested in the generalization loss $\mathscr{E}(\boldsymbol{\theta};\mathcal{L}) := \mathbb{E}_{(\mathrm{x},\mathrm{y})}[\mathcal{L}(\mathrm{x},\mathrm{y},\boldsymbol{\theta})]$. The classic way to propose an estimator iteratively is SGD,

[2] note this implies L-smoothness

for which it is well known that SGD is the discretized gradient flow (i.e. a "discrete derivative") with respect to the loss function. Since we train with random iterates, the sequence $\left(\theta_k^{(\gamma)}\right)_{k\in\mathbb{N}}$ will be random itself. We provide more details in appendix C.

2.III *From general to specific and back*

We are now ready to present the main statements. We will not prove all of them but give intuition. The common observation is that all results attempt to take a clever asymptotic/limit. To begin, we show that the object of interest is well-defined. Throughout, we will take the assumptions as granted (let all of them hold in each statement for proper parameters). For further details, we refer to the original text.

Proposition 2.2 (Prop.2 in (Dieuleveut, Durmus, and Bach 2020))**.** *Let $\gamma \in (0, 2/L)$. The iterations of SGD seen as a Markov chain admit a unique stationary distribution π_γ that has finite second moments, i.e. $\pi_\gamma \in \mathcal{P}_2(\mathbb{R}^d)$. We can also **quantify** the rate of convergence in two ways: via the Wasserstein distance over probability measures (def. D.16) and in terms of "nice" functions integrated out by the kernel over time. Mathematically for all $\boldsymbol{\theta} \in \mathbb{R}^d$ and $k \in \mathbb{N}$:*

initial distribution — rate function; goes exp fast to zero — invariant measure — mean square distance in invariant measure (initial conditions)

$$\mathrm{W}_2^2\left(\, R_\gamma^k(\boldsymbol{\theta}, \cdot)\ ,\ \pi_\gamma\ \right) \leqslant [\hbar(L,\mu,\gamma)]^k \int_{\mathbb{R}^d} \|\boldsymbol{\theta} - \boldsymbol{\xi}\|_2^2 \,\mathrm{d}\pi_\gamma(\boldsymbol{\xi})\ , \tag{2.3}$$

and for all $\boldsymbol{\theta} \in \mathbb{R}^d$, $k \in \mathbb{N}$ and $\phi : \mathbb{R}^d \to \mathbb{R}$ Lipschitz with constant L_ϕ:

$$\left| R_\gamma^k \phi(\boldsymbol{\theta}) - \pi_\gamma(\phi) \right| \leqslant L_\phi [\hbar(L,\mu,\gamma)]^{k/2} \sqrt{\int_{\mathbb{R}^d} \|\boldsymbol{\theta} - \boldsymbol{\xi}\|_2^2 \,\mathrm{d}\pi_\gamma(\boldsymbol{\xi})}. \tag{2.4}$$

Naturally, the function $\hbar \equiv \hbar(L,\mu,\gamma,k)$ is a function that is less than one exactly because we have the stability condition on the step-size.

Proof. See appendix B. □

Since we have a stationary distribution, it makes sense to see how the statistics behave when we start already at the stationary distribution, and when we reach it from another. This is the main lesson of the next two statements.

Corollary 2.5 (Prop. 16 in (Dieuleveut, Durmus, and Bach 2020))**.** *The running mean converges to the stationary mean at rate $O(1/k)$.*

Proposition 2.6 (Prop. 3 and Thm. 4 in (Dieuleveut, Durmus, and Bach 2020), for $\boldsymbol{\theta}_0 \sim \pi_\gamma$ already stationary)**.** *Consider first our baby example. Let $\boldsymbol{\Sigma} = \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]$ be positive definite and $\gamma \in (0, 2/L)$. Then we have:*

- $\overline{\boldsymbol{\theta}}_\gamma = \boldsymbol{\theta}^\star$, *so the mean of the stationary distribution is aligned with the optimum;*
- *the deviations are error dependent:*

$$\int_{\mathbb{R}^d} [\boldsymbol{\theta} - \boldsymbol{\theta}^\star]\,[\boldsymbol{\theta} - \boldsymbol{\theta}^\star]^\top \,\mathrm{d}\pi_\gamma(\boldsymbol{\theta}) = \gamma \boldsymbol{R}(\boldsymbol{\Sigma}, \mathbb{P}_{\boldsymbol{\epsilon}(\theta)}, \pi_\gamma), \tag{2.7}$$

  where $\boldsymbol{R}$ is explicit. Moreover, for this easy case we can find a nice expression in terms of the multiplicative noise, see appendix B.

In general, we find that the distribution is not guessing the optimal point right, and we will have:

$$\overline{\boldsymbol{\theta}}_\gamma - \boldsymbol{\theta}^\star = \gamma \boldsymbol{g}(f, \boldsymbol{\theta}^\star) + O\left(\gamma^2\right), \qquad \int_{\mathbb{R}^d} [\boldsymbol{\theta} - \boldsymbol{\theta}^\star][\boldsymbol{\theta} - \boldsymbol{\theta}^\star]^\top \,\mathrm{d}\pi_\gamma(\boldsymbol{\theta}) = \gamma \boldsymbol{P}(f, \boldsymbol{\theta}^\star) + O\left(\gamma^2\right). \tag{2.8}$$

Again, $\boldsymbol{g}, \boldsymbol{P}$ are explicit.

Remark 2.9. *The distance mean-optimum and the oscillations are of order $\gamma, \sqrt{\gamma}$ respectively.*

Theorem 2.10 (Cor. 6 and Thm. 5 in (Dieuleveut, Durmus, and Bach 2020), for generic $\boldsymbol{\theta}_0$)**.** *For our baby example, we have quantitative CLT when γ is small enough*[3] *$\forall \boldsymbol{\theta}_0 \in \mathbb{R}^d$:*

$$\mathbb{E}\left[\overline{\boldsymbol{\theta}}_k^{(\gamma)}\right] - \boldsymbol{\theta}^\star = \frac{1}{k\gamma}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star) + O\left(\hbar^k\right) \tag{2.11}$$

$$\mathbb{E}\left[[\overline{\boldsymbol{\theta}}_k^{(\gamma)} - \overline{\boldsymbol{\theta}}_\gamma][\overline{\boldsymbol{\theta}}_k^{(\gamma)} - \overline{\boldsymbol{\theta}}_\gamma]^\top\right] = \frac{1}{k}\boldsymbol{U}(\boldsymbol{\Sigma}, \mathbb{P}_{\boldsymbol{\epsilon}(\boldsymbol{\theta})}, \pi_\gamma) + \frac{1}{k^2\gamma^2}\boldsymbol{Q}(\boldsymbol{\theta}_0, \boldsymbol{\theta}^\star, \boldsymbol{\Sigma}, \pi_\gamma, \mathbb{P}_{\mathbf{x}}) + \frac{1}{k^2\gamma^2}\boldsymbol{S}(\boldsymbol{\theta}^\star, \boldsymbol{\Sigma}, \pi_\gamma) + O\left(k^3\right). \tag{2.12}$$

The bold matrices are explicit. In the general case, the formulas are expressed implicitly in terms of ***Poisson solution*** *of functions, but exhibit the same decomposition.*

Remark 2.13. *If we start $\theta_0 \sim \pi_\gamma$, then the variance term is zero in the baby example!*

Lastly, we make use of a nice interpretation of SGD as a discretized gradient flow. It will allow us to eventually relate the *actual chain* to the optimal value.

Theorem 2.14 (Thm. 7 in (Dieuleveut, Durmus, and Bach 2020))**.** *Under appropriate additional assumptions, a nice class of functions g satisfies for any $\boldsymbol{\theta}_0 \in \mathbb{R}^d$:*

$$\mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k} g\left(\theta_i^{(\gamma)}\right) - g(\boldsymbol{\theta}^\star)\right] = \frac{1}{k\gamma}\mathfrak{v}(\boldsymbol{\theta}_0, k+1, \gamma) + \frac{\gamma}{2}\mathfrak{b}(\boldsymbol{\theta}^\star, \mathbb{P}_{\boldsymbol{\epsilon}}, g) - \frac{\gamma}{k}A_1(\boldsymbol{\theta}_0) - \gamma^2 A_2(\boldsymbol{\theta}_0, k), \tag{2.15}$$

$$A_1(\boldsymbol{\theta}_0) \leqslant C\left(1 + \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star\|_2^{\tilde{p}}\right), \qquad A_2(\boldsymbol{\theta}_0, k) \leqslant C\left(1 + \frac{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star\|_2^{\tilde{p}}}{k}\right). \tag{2.16}$$

Here $C > 0, \tilde{p}$ are constants, the latter being explicit and $\mathfrak{v}, \mathfrak{b}$ are implicit scalars.

Remark 2.17. *In particular, there exists $C_1, C_2(\boldsymbol{\theta}_0) \geqslant 0$ such that:*

$$\mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k}\left\|\theta_i^{(\gamma)} - \boldsymbol{\theta}^\star\right\|_2^{2q}\right] = C_1\gamma + \frac{1}{k}C_2(\boldsymbol{\theta}_0) + O\left(\gamma^2\right). \tag{2.18}$$

Takeaway

Under appropriate assumptions:

(R1) initial conditions are forgotten exponentially fast and any chain converges to a unique distribution;

(R2) there is $\boldsymbol{d}$ such that for small enough step-size:

$$\overline{\boldsymbol{\theta}}_\gamma = \int_{\mathbb{R}^d} \xi \mathrm{d}\pi_\gamma(\xi) = \boldsymbol{\theta}^\star + \gamma\boldsymbol{d} + \boldsymbol{r}_\gamma^{(1)}, \qquad \left\|\boldsymbol{r}_\gamma^{(1)}\right\| \leqslant C\gamma^2; \tag{2.19}$$

(R3) the bias is expanded as:

$$\mathbb{E}\left[\overline{\theta}_k^{(\gamma)} - \boldsymbol{\theta}^\star\right] = \frac{A(\boldsymbol{\theta}_0, \gamma)}{k} + \gamma\boldsymbol{d} + \boldsymbol{r}_\gamma^{(2)}, \qquad \left\|\boldsymbol{r}_\gamma^{(2)}\right\|_2 \leqslant C(\gamma^2 + \mathrm{e}^{-k\mu\gamma}); \tag{2.20}$$

(R4) there is a quantitative CLT for the variance terms for fixed γ as $k \to \infty$:

$$\mathbb{E}\left[\left\|\overline{\theta}_k^{(\gamma)} - \overline{\boldsymbol{\theta}}_\gamma\right\|^2\right] = B_1(\gamma)\frac{1}{k} + B_2(\gamma)\frac{1}{k^2} + O\left(\frac{1}{k^3}\right). \tag{2.21}$$

Combining variance and bias, we have characterized the trade-off in SGD.

Remark 2.22. *Like the Euler-Maruyama scheme the weak error expansion of SGD in step-size γ between π_γ and $\delta_{\boldsymbol{\theta}^\star}$ is of order $O(\gamma)$ (Talay and Tubaro 1990).*

[3] Needs to satisfy a bound wrt to inflated empirically covariances and true covariances of $\mathbb{P}_{\mathbf{x}}$, see (Dieuleveut, Durmus, and Bach 2020, eqn. 9)

We provide an overview of two interesting and less traditional aspects of the work. One is numerically oriented and quick to present, the other is rather deep.

RICHARDSON-ROMBERG Result (R2) above suggests using Richardson-Romberg extrapolation to decrease the bias of iterations. The idea is quick to understand. Recall that we have the weak error expansion between the stationary integral and the Dirac indicator at the true minimum, that is:

$$\int_{\mathbb{R}^d} g(\boldsymbol{\xi})\,\mathrm{d}\pi_\gamma(\boldsymbol{\xi}) = g(\boldsymbol{\theta}^\star) + \gamma C_1^g + \boldsymbol{r}_\gamma^g, \qquad \left\|\boldsymbol{r}_\gamma^g\right\|_2 \leqslant C_2^g\gamma^2, \qquad \text{for all } g \text{ "nice".} \tag{3.1}$$

Therefore, if we run two chains at $\gamma, 2\gamma$ step-size, we will have that $\overline{\theta}_k^{(\gamma)} \to \overline{\boldsymbol{\theta}}_\gamma, \overline{\theta}_k^{(2\gamma)} \to \overline{\boldsymbol{\theta}}_{2\gamma}$ with distance to the optimal given by (choose $g = \mathrm{Id}$ the identity map):

$$\overline{\boldsymbol{\theta}}_\gamma = \boldsymbol{\theta}^\star + \gamma \boldsymbol{d}_1^{\mathrm{Id}} + \boldsymbol{r}_\gamma^{\mathrm{Id}}, \qquad \overline{\boldsymbol{\theta}}_\gamma = \boldsymbol{\theta}^\star + 2\gamma \boldsymbol{d}_1^{\mathrm{Id}} + \boldsymbol{r}_{2\gamma}^{\mathrm{Id}}, \qquad \max\left\{\left\|2\boldsymbol{r}_\gamma^{\mathrm{Id}}\right\|, \left\|\boldsymbol{r}_{2\gamma}^{\mathrm{Id}}\right\|\right\} \leqslant 2C\gamma^2. \tag{3.2}$$

Taking the combined chain running means: $\left(2\overline{\theta}_k^{(\gamma)} - \overline{\theta}_k^{(2\gamma)}\right)_{k\in\mathbb{N}}$ will then cancel out the linear $O(\gamma)$ term and improve the distance to $\boldsymbol{\theta}^\star$ up to factors of $O(\gamma^2)$.

Remark 3.3. *The cost is an increase in the variance which stays of the same order.*

WHAT IS A POISSON SOLUTION? Lastly, we try to give some intuition on generators and Poisson solutions. If we rescale time, we can see equation eq. (1.1) as a noisy gradient flow, which would be just $\dot{\boldsymbol{\theta}}_t = -\nabla f(\boldsymbol{\theta}_t)$. We could look at a set of functions that may represent a meaningful derivative of the flow of points. Denoting such flow as $\varphi_t(\boldsymbol{\theta})$, we define the infinitesimal generator $\mathcal{A}$ as:

$$\mathcal{A}h(\boldsymbol{\theta}) := \lim_{t\downarrow 0} \frac{1}{t}\left(h(\varphi_t(\boldsymbol{\theta})) - h(\boldsymbol{\theta})\right), \tag{3.4}$$

where we apply functions $h : \mathbb{R}^d \to \mathbb{R}$ that are such that the limit exists. For simplicity, we will say that $\mathcal{A}$ has domain $\mathrm{D}(\mathcal{A})$ with these nice functions. From the definition of Markov property, we have that $\mathbb{E}\left[h(\theta_{k+1}) \mid \mathcal{F}_k\right] = R_\gamma h(\theta_k)$, and by the construction proposed we will also have the approximate relation:

$$\mathbb{E}\left[h(\theta_{t+s}) \mid \mathcal{F}_t\right] \approx h(\theta_t) + t\mathcal{A}h(\theta_t). \tag{3.5}$$

With this in hand, one can show that $\mathrm{d}/\mathrm{d}t h(\varphi_t(\boldsymbol{\theta})) = \mathcal{A}\varphi_t(\boldsymbol{\theta})$ and that we can make the interpretation of a derivative in time to conclude that $\varphi_t(\theta) = u(t, \boldsymbol{\theta})$ is a solution to the PDE:

$$\frac{\partial}{\partial t} u(t, \boldsymbol{\theta}) = \mathcal{A}u(t, \boldsymbol{\theta}). \tag{3.6}$$

In a somewhat more interesting perspective, from these observations it can be shown that:

$$\mathrm{m}_t^h := h(\theta_t) - h(\theta_0) - \int_0^t \mathcal{A}h(\theta_s)\,\mathrm{d}s, \tag{3.7}$$

is a martingale, and thus obeys a central limit theorem. Let us see this through a quicker example that removes many details. Under our assumptions certain Poisson equations are well-behaved and there is a more direct link to at least an asymptotic form of the CLT. Given a kernel K_γ with a unique invariant distribution and a function $h \in \mathrm{L}^1(\pi)$, we reorder the random sum:

$$\mathrm{s}_n(h) := \sum_{k=0}^{n-1} h(\theta_k) = h(\theta_k) - R_\gamma h(\theta_k) + R_\gamma h(\theta_k) + h(\theta_{k-1}) - R_\gamma^2 h(\theta_k) - R_\gamma h(\theta_{k-1}) + \cdots, \tag{3.8}$$

stressing that we might decompose each iteration steps into martingales. In this spirit, we say $\hat{h}$ is a solution to the Poisson equation if (Douc et al. 2018, chaps. 22-23):

$$\hat{h} - R_\gamma \hat{h} = h - \pi(h), \qquad \pi\text{-a.e., } h \in \mathrm{L}^1(R_\gamma). \tag{3.9}$$

The intuitive interpretation is as follows. The transition of the law of a Markov chain can be seen through the lenses of the (functional) heat equation $\partial_t \hat{h} = (I - K_\gamma)\hat{h} = \Delta\hat{h}$. If $\Delta\hat{h} = 0$, the function is said to be harmonic, while if $\Delta\hat{h} = 0$ for all $\hat{h}$, the chain is trivially stationary. In this case, we give a "derivative" h and a kernel K_γ and want to find which function we were actually working on. A trivial substitution of the identity shows us that:

$$\mathrm{s}_n(h) = \hat{h}(\theta_0) - \hat{h}(\theta_n) + \underbrace{\sum_{k=1}^{n} \hat{h}(\theta_k) - R_\gamma \hat{h}(\theta_{k-1})}_{\mathrm{m}_n^h}, \tag{3.10}$$

where m_n^h is a martingale.[4] The normalization by $1/\sqrt{n}$ and the CLT for martingales comes into rescue (Douc et al. 2018, chaps. 21-22). In appendix D we report a continuous time generalization. Now the problem is quantifying this convergence.

4 DISCUSSION

While we have only scratched the surface, it is possible to see that under our rather restrictive assumption the theory is basically closed, as we have a full characterization of the stochastic nature of the phenomenon that is non-asymptotic. As common in the theory of optimization, going beyond the combination of L-smooth and μ-strong convexity while retaining an explicit convergence rate is hard. One potential expansion in this direction could be modelling the quadratic nature of the function locally, at the cost of largely complicating the expressions. A potentially related theoretical principle is that of the resolvent method, which expands the concept of Poisson equation to a larger class of functions (see Douc et al. (2018, chaps. 21-22)). The real obstacle being the objective function, one could also slightly decrease the generality of the results and still inspect SGD under different function classes (as was done over the years), but still apply the generic tools present. In this regard, it is understandable that the result of uniqueness of the invariant distribution is striking and useful, therefore it would be crucial to see if it can be kept, or it has to be given up.
Lastly, there is a renewed interest on one-pass SGD for high-dimensional problems and its interplay with hardness of inference. There, the attempt is to classify functions based on the computational time needed to minimize them. While the setting is very different, Collins-Woodfin et al. (2023) use a resolvent method to transform a very implicit ODE into a solvable complex ODE, and express their solutions as a set of implicit equations.

REFERENCES

Aguech, Rafik, Eric Moulines, and Pierre Priouret (Jan. 2000). "On a Perturbation Approach for the Analysis of Stochastic Tracking Algorithms". In: *SIAM Journal on Control and Optimization* 39.3, pp. 872–899. ISSN: 0363-0129. DOI: 10.1137/S0363012998333852. (Visited on 12/11/2024).

Chen, Xi et al. (Feb. 2020). "Statistical Inference for Model Parameters in Stochastic Gradient Descent". In: *The Annals of Statistics* 48.1, pp. 251–273. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/18-AOS1801. (Visited on 12/11/2024).

Collins-Woodfin, Elizabeth et al. (2023). *Hitting the High-Dimensional Notes: An ODE for SGD learning dynamics on GLMs and multi-index models*. arXiv: 2308.08977 [math.OC]. URL: https://arxiv.org/abs/2308.08977.

Defossez, Alexandre and Francis Bach (Feb. 2015). "Averaged Least-Mean-Squares: Bias-Variance Trade-offs and Optimal Sampling Distributions". In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 205–213. (Visited on 12/11/2024).

Dieuleveut, Aymeric, Alain Durmus, and Francis Bach (June 2020). "Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains". In: *The Annals of Statistics* 48.3, pp. 1348–1382. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/19-AOS1850. (Visited on 12/11/2024).

Dieuleveut, Aymeric, Nicolas Flammarion, and Francis Bach (2017). "Harder, Better, Faster, Stronger Convergence Rates for Least-Squares Regression". In: *Journal of Machine Learning Research* 18.101, pp. 1–51. ISSN: 1533-7928. (Visited on 12/11/2024).

Douc, Randal et al. (2018). *Markov Chains*. Springer Series in Operations Research and Financial Engineering. Cham: Springer International Publishing. ISBN: 978-3-319-97703-4 978-3-319-97704-1. DOI: 10.1007/978-3-319-97704-1. (Visited on 12/11/2024).

[4] Just verify that $\mathbb{E}\left[\mathrm{m}_n^h \mid \mathcal{F}_{n-1}\right] - \mathrm{m}_{n-1}^h = \mathbb{E}\left[\hat{h}(\theta_n) \mid \mathcal{F}_{n-1}\right] - R_\gamma \hat{h}(\theta_{k-1}) = 0$.

Fort, Jean-Claude and Gilles Pagès (Jan. 1999). "Asymptotic Behavior of a Markovian Stochastic Algorithm with Constant Step". In: *SIAM Journal on Control and Optimization* 37.5, pp. 1456–1482. ISSN: 0363-0129. DOI: 10.1137/S0363012997328610. (Visited on 12/11/2024).

Le Gall, Jean-François (2016). "General Theory of Markov Processes". In: *Brownian Motion, Martingales, and Stochastic Calculus*. Ed. by Jean-François Le Gall. Cham: Springer International Publishing, pp. 151–184. ISBN: 978-3-319-31089-3. DOI: 10.1007/978-3-319-31089-3_6. (Visited on 12/16/2024).

Ljung, L., G. Pflug, and H. Walk (Dec. 2012). *Stochastic Approximation and Optimization of Random Systems*. Birkhäuser. ISBN: 978-3-0348-8609-3.

Su, Weijie J. and Yuancheng Zhu (Aug. 2018). *Uncertainty Quantification for Online Learning and Stochastic Approximation via Hierarchical Incremental Gradient Descent*. DOI: 10.48550/arXiv.1802.04876. arXiv: 1802.04876 [stat]. (Visited on 12/11/2024).

Talay, Denis and Luciano Tubaro (Jan. 1990). "Expansion of the Global Error for Numerical Schemes Solving Stochastic Differential Equations". In: *Stochastic Analysis and Applications* 8.4, pp. 483–509. ISSN: 0736-2994. DOI: 10.1080/07362999008809220. (Visited on 12/11/2024).

Villani, Cédric (2009). *Optimal Transport*. Ed. by M. Berger et al. Vol. 338. Grundlehren Der Mathematischen Wissenschaften. Berlin, Heidelberg: Springer. ISBN: 978-3-540-71049-3 978-3-540-71050-9. DOI: 10.1007/978-3-540-71050-9. (Visited on 12/11/2024).

A ASSUMPTIONS

In this section, we rewrite the assumptions for completeness.

Assumption A.1. *The function f is μ-strongly convex, in the sense of definition D.1.*

Assumption A.2. *The functions f is five times differentiable with continuous and uniformly bounded derivatives. In particular, it is L-smooth.*

Assumption A.3. *There is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$ for which:*

- *$\boldsymbol{\epsilon}_k(\boldsymbol{\theta})$ is adapted for all $\boldsymbol{\theta} \in \mathbb{R}^d$;*
- *$\mathbb{E}\left[\boldsymbol{\epsilon}_{k+1}(\boldsymbol{\theta}) \mid \mathcal{F}_k\right] = 0$ for all $\boldsymbol{\theta} \in \mathbb{R}^d$;*
- *$(\boldsymbol{\epsilon}_k)_{k\in\mathbb{N}}$ are i.i.d. random $\mathbb{R}^d \mapsto \mathbb{R}^d$ functions (i.e. random fields);*[5]
- *$\boldsymbol{\theta}_0 \in \mathcal{F}_0$.*

Assumption A.4. *This condition depends on some p to be chosen. The function f_k is almost surely C-co-coercive (def. D.13) with $C = L$ the smoothness constant. For the error, the p norm at the optimum is controlled as:*

$$\left(\mathbb{E}\left[\|\boldsymbol{\epsilon}_k(\boldsymbol{\theta}^\star)\|\right]_2^p\right)^{\frac{1}{p}} \leqslant \tau_p, \qquad \text{for some } \tau_p.$$

Assumption A.5. *The function $\boldsymbol{C}(\boldsymbol{\theta}) := \mathbb{E}\left[[\boldsymbol{\epsilon}(\boldsymbol{\theta})\boldsymbol{\epsilon}(\boldsymbol{\theta})]^\top\right]$ is in $\mathrm{C}^3(\mathbb{R}^d, \mathbb{R}^{d\times d})$ and such that:*

$$\max_{i\in\{1,2,3\}} \left\|\boldsymbol{C}^{(i)}(\boldsymbol{\theta})\right\| \leqslant M_\epsilon \left(1 + \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^{k_\epsilon}\right), \qquad \text{for some } k_\epsilon, M_\epsilon, \text{for all } \boldsymbol{\theta} \in \mathbb{R}^d,$$

where $\mathbf{C}^{(i)}$ are the (tensor) derivatives of the matrix $\boldsymbol{C}$.

Assumption A.6. *This condition depends on the tuple (ℓ, p). Let g be polynomially locally Lipschitz, in the sense of definition D.11. There are positive a_g, b_g such that $g \in \mathrm{C}^\ell(\mathbb{R}^d)$ and:*

$$\left\|g^{(i)}(\boldsymbol{\theta})\right\| \leqslant a_g \left(b_g + \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^p\right), \qquad \forall \boldsymbol{\theta} \in \mathbb{R}^d, i \in \{1, \dots, \ell\}, \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d,$$

where $g^{(i)}$ are the i^{th} derivatives of g, which become tensors for $i > 2$.

Assumption A.7 (A.4 bis)**.** *Alternatively, suppose:*

- *for some $\tilde{\tau}_p \geqslant 0$ it holds that $\left(\mathbb{E}\left[\|\boldsymbol{\epsilon}(\boldsymbol{\theta})\|_2^p\right]\right)^{1/p} \leqslant \tilde{\tau}_p$;*
- *the smoothness constant L is such that:*

$$\mathbb{E}\left[\|\nabla \mathrm{f}_1(\boldsymbol{x}) - \nabla \mathrm{f}_1(\boldsymbol{y})\|_2^q\right] \leqslant L^{q-1}\|\boldsymbol{x} - \boldsymbol{y}\|_2^{q-2}\langle \boldsymbol{x} - \boldsymbol{y}, \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\rangle, \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d, \forall q \in \{2, \dots, p\}.$$

  where we write f_1 for the random observed gradient at the 1^{st} iteration (they are i.i.d. so it does not matter).

Assumption A.8 (A.4 tris)**.** *Alternatively, suppose there exists a **global** $\tau \geqslant 0$ for which $\sup_{\boldsymbol{\theta}\in\mathbb{R}^d}\left(\mathbb{E}\left[\|\boldsymbol{\epsilon}(\boldsymbol{\theta})\|_2^4\right]\right)^{1/4} \leqslant \tau$.*

A.1 *Remarks on assumptions*

The assumption of random fields is weaker than assuming $(\boldsymbol{\epsilon}_k(\boldsymbol{\theta}))_{k\in\mathbb{N}}$ is i.i.d. for all $\boldsymbol{\theta}$, since it is not at fixed $\boldsymbol{\theta}$ but globally.
Assumption A.6 is only needed for theorem 2.14.
The reason to introduce Assumptions A.7-A.8 it to make clear how to avoid the bounded assumption of A.4. In particular, assumption A.7 is the weakest, and assumption A.8 is the strongest. Moreover, if we make the assumption of i.i.d. errors as above, even just for the sequence $(\boldsymbol{\epsilon}_k(\boldsymbol{\theta}_{k-1}))_{k\in\mathbb{N}}$ is sufficient to imply assumption A.8 alone. In most works, the noise is taken to be "completely independent" as in this example, which is technically termed **semi-stochastic**. In such regard, this work is far more general than classical ML analysis.
Let us now make clear which is needed for which.

[5] This condition is global: we see $\boldsymbol{\epsilon}$ as a random mapping $\mathbb{R}^d \ni \boldsymbol{\theta} \mapsto \boldsymbol{\epsilon}(\boldsymbol{\theta}) \in \mathbb{R}^d$, so we do not state $\forall \boldsymbol{\theta} \in \mathbb{R}^d$ merely because it is not fixed!

- For proposition 2.2 we need A.1-A.2-A.3-A.4 with $p = 2$;
- for proposition 2.6 we need A.1-A.2-A.3-A.4, with $p = 4$ in the special case and $p = \max\{6, 2(k_\epsilon + 1)\}$ in general;
- for theorem 2.10 we need A.1-A.2-A.3-A.4 with $p = 4$;
- for theorem 2.14 we need A.1-A.2-A.4-A.5-A.6 with:
  - an additional condition on the noise, such that for some $q \in \mathbb{N}, C \geqslant 0$ for all $\boldsymbol{\theta} \in \mathbb{R}^d$ one has:

$$\mathbb{E}\left[\|\epsilon_1(\boldsymbol{\theta})\|_2^{p+k_\epsilon}\right] \leqslant C(1 + \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^q); \tag{A.9}$$

  - A.6 holding with $\ell = 5, p$, i.e. the five times continuous differentiability and the p from the other assumptions match;
  - A.4 holding for a slightly tweaked $\tilde{p} = \max\{p + 3 + q, k_\epsilon\}$;
  - the step size being changed to $\gamma \in (0, 1/\zeta L)$ for $\zeta \equiv \zeta(\tilde{p}) > 0$.

It is clear that the last result is more technical than the others.

B SOME PROOFS AND OTHER STATEMENTS

Let us recall the full statement to have it ready.

Proposition 2.2 (Prop.2 in (Dieuleveut, Durmus, and Bach 2020)). *Let $\gamma \in (0, 2/L)$. The iterations of SGD seen as a Markov chain admit a unique stationary distribution π_γ that has finite second moments, i.e. $\pi_\gamma \in \mathcal{P}_2(\mathbb{R}^d)$. We can also **quantify** the rate of convergence in two ways: via the Wasserstein distance over probability measures (def. D.16) and in terms of "nice" functions integrated out by the kernel over time. Mathematically for all $\boldsymbol{\theta} \in \mathbb{R}^d$ and $k \in \mathbb{N}$:*

$$W_2^2(\, R_\gamma^k(\boldsymbol{\theta}, \cdot)\,,\ \pi_\gamma\,) \leqslant [\hbar(L, \mu, \gamma)]^k \int_{\mathbb{R}^d} \|\boldsymbol{\theta} - \boldsymbol{\xi}\|_2^2 \,\mathrm{d}\pi_\gamma(\boldsymbol{\xi})\,, \tag{2.2}$$

initial distribution; rate function; goes exp fast to zero; invariant measure; mean square distance in invariant measure (initial conditions)

and for all $\boldsymbol{\theta} \in \mathbb{R}^d$, $k \in \mathbb{N}$ and $\phi : \mathbb{R}^d \to \mathbb{R}$ Lipschitz with constant L_ϕ:

$$\left|R_\gamma^k \phi(\boldsymbol{\theta}) - \pi_\gamma(\phi)\right| \leqslant L_\phi [\hbar(L, \mu, \gamma)]^{k/2} \sqrt{\int_{\mathbb{R}^d} \|\boldsymbol{\theta} - \boldsymbol{\xi}\|_2^2 \,\mathrm{d}\pi_\gamma(\boldsymbol{\xi})}. \tag{2.3}$$

Naturally, the function $\hbar \equiv \hbar(L, \mu, \gamma, k)$ is a function that is less than one exactly because we have the stability condition on the step-size.

Proof. The idea of the proof is very simple: we know that the Wasserstein distance is very nice: namely it makes $\mathcal{P}_2(\mathbb{R}^d)$ complete, therefore reducing the question of convergence to proving that the sequence is Cauchy. In our case, we have a sequence of probability measures generated by multiple applications of the same kernel R_γ, therefore reducing the question to possibly bounding a single kernel.[6] Then, we have plenty of assumptions on the function and the type of randomness we allow, which will hopefully bring us from a rather abstract distance over probability measures to a friendlier Euclidean distance, that is also crucially at one step before. As we will see, finding a direct bound of the Wasserstein distance in terms of Euclidean distance is by a routine argument, but the assumptions on our problem are really what makes it possible to come back by one-step while keeping the inequality. Once we forget one step, by the fact that again we apply the same kernel, we will be able to iterate infinitely these steps. Hopefully, the multiplied discount at each iteration will be fast enough to establish convergence. Once this fact is established, existence and uniqueness of the invariant probability measure follow trivially by special choices of our more general master equation of

[6] Notice that here it is crucial to have the same step-size to make the reasoning this simple.

"fast convergence".
We enforce the stability condition that $\gamma \in (0, 2/L)$, and consider two starting distributions $\lambda_1, \lambda_2 \in \mathcal{P}_2(\mathbb{R}^d)$. A standard result in optimal transport (see Villani (2009, thm. 4.1)) states that the Wasserstein distance can be represented as a distance of random variables $\theta_0^{(1)}, \theta_0^{(2)}$ independent of the noise field $(\epsilon_k)_{k\in\mathbb{N}}$. In other words, we have $W_2^2(\lambda_1, \lambda_2) = \mathbb{E}\left[\left\|\theta_0^{(1)} - \theta_0^{(2)}\right\|_2^2\right]$. Let us run these two chains with the same step size γ, omitted in the notation for simplicity. Let us run the two chains with the same noise field, so that their gradients are respectively:

$$\nabla f_k(\theta_k^{(1)}) = \nabla f(\theta_k^{(1)}) + \epsilon_{k+1}(\theta_k^{(1)}); \tag{B.1}$$
$$\nabla f_k(\theta_k^{(2)}) = \nabla f(\theta_k^{(2)}) + \epsilon_{k+1}(\theta_k^{(2)}). \tag{B.2}$$

Let us use the independence of chain and noise to show that noise is de-correlated across chains. As a canonical example we have:

$$\mathbb{E}\left[\left\langle \theta_0^{(1)}, \epsilon(\theta_0^{(2)})\right\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\left\langle \theta_0^{(1)}, \epsilon(\theta_0^{(2)})\right\rangle\right] \mid \mathcal{F}_0\right] \overset{A.3}{=} \mathbb{E}\left[\left\langle \theta_0^{(1)}, \mathbb{E}\left[\epsilon(\theta_0^{(2)}) \mid \mathcal{F}_0\right]\right\rangle\right] \overset{A.3}{=} 0. \tag{B.3}$$

By symmetry, the result will hold if we swap the roles. A moment of thought also shows that the result holds if we consider the chain and the noise of the same chain, or any measurable function obviously. By construction, the combined chain distribution belongs to the set of couplings $\Pi(\lambda R_\gamma^k, \lambda_2 R_\gamma^k)$ and since the Wasserstein distance is an infimum, we can write a one-step inequality:

$$W_2^2(\lambda_1 R_\gamma, \lambda_2 R_\gamma) \leqslant \mathbb{E}\left[\left\|\theta_1^{(1)} - \theta_1^{(2)}\right\|_2^2\right] \tag{B.4}$$
$$= \mathbb{E}\left[\left\|\theta_0^{(1)} - \gamma\nabla f_1(\theta_0^{(1)}) - \theta_0^{(2)} + \gamma\nabla f_1(\theta_0^{(1)})\right\|_2^2\right], \tag{B.5}$$
$$= \mathbb{E}\left[\left\|\theta_0^{(1)} - \gamma\nabla f(\theta_0^{(1)}) - \gamma\epsilon_1(\theta_0^{(1)}) - \theta_0^{(2)} + \gamma\nabla f(\theta_0^{(1)}) + \gamma\epsilon_1(\theta_0^{(1)})\right\|_2^2\right], \tag{B.6}$$

where unrolled the step of the recursion. It is now natural to seek to exploit the orthogonality we just found. Let us then square the difference of starting points minus the difference of stochastic gradients. We find:

$$\mathbb{E}\left[\left\|\theta_1^{(1)} - \theta_1^{(2)}\right\|_2^2\right] = \mathbb{E}\left[\left\|\theta_0^{(1)} - \theta_0^{(2)}\right\|_2^2\right] + \gamma^2\mathbb{E}\left[\left\|\nabla f_1(\theta_0^{(1)}) - \nabla f_2(\theta_0^{(2)})\right\|_2^2\right] \tag{B.7}$$
$$- 2\gamma\mathbb{E}\left[\left\langle \underbrace{\nabla f_1(\theta_0^{(1)}) - \nabla f_1(\theta_0^{(2)})}_{=\nabla f(\theta_0^{(1)}) + \epsilon_1(\theta_0^{(1)}) - \nabla f(\theta_0^{(2)}) - \epsilon_1(\theta_0^{(2)})}, \theta_0^{(1)} - \theta_0^{(2)}\right\rangle\right]. \tag{B.8}$$

The second term is not really friendly, but we can use our result quite directly by cancelling the noise dependent terms! We therefore find a more amenable expression:

$$\mathbb{E}\left[\left\|\theta_1^{(1)} - \theta_1^{(2)}\right\|_2^2\right] = \mathbb{E}\left[\left\|\theta_0^{(1)} - \theta_0^{(2)}\right\|_2^2\right] + \gamma^2\mathbb{E}\left[\left\|\nabla f_1(\theta_0^{(1)}) - \nabla f_2(\theta_0^{(2)})\right\|_2^2\right] \tag{B.9}$$
$$- 2\gamma\mathbb{E}\left[\left\langle \nabla f(\theta_0^{(1)}) - \nabla f(\theta_0^{(2)}), \theta_0^{(1)} - \theta_0^{(2)}\right\rangle\right]. \tag{B.10}$$

Let us stress that inside the norm we will have the *noisy* gradient while in the inner product we have the *unbiased* gradient evaluated at the random iterates. At this moment we need to exploit the assumptions on the function class we chose to get rid of the random gradients. By L-co-coercivity of f_1 (def. D.13) assumed in A.4 we find that:

$$\mathbb{E}\left[\left\|\nabla f_1(\theta_0^{(1)}) - \nabla f_2(\theta_0^{(2)})\right\|_2^2\right] \leqslant L\mathbb{E}\left[\left\langle \nabla f_1(\theta_0^{(1)}) - \nabla f_2(\theta_0^{(2)}), \theta_0^{(1)} - \theta_0^{(2)}\right\rangle\right], \tag{B.11}$$

and another application of independence allows us to say that the noise cancels with the vectors in the expectation of the inner product. Thus,

$$\mathbb{E}\left[\left\|\theta_1^{(1)} - \theta_1^{(2)}\right\|_2^2\right] \leqslant \mathbb{E}\left[\left\|\theta_0^{(1)} - \theta_0^{(2)}\right\|_2^2\right] - 2\gamma\left(1 - \gamma\frac{L}{2}\right)\mathbb{E}\left[\left\langle \nabla f(\theta_0^{(1)}) - \nabla f(\theta_0^{(2)}), \theta_0^{(1)} - \theta_0^{(2)}\right\rangle\right]. \tag{B.12}$$

Now that we have recovered an expression that depends only on f, we can use the strong convexity from assumption A.1 to conclude:

$$\mathbb{E}\left[\left\|\theta_1^{(1)} - \theta_1^{(2)}\right\|_2^2\right] \leqslant \left[1 - 2\mu\gamma\left(1 - \gamma\frac{L}{2}\right)\right]\mathbb{E}\left[\left\|\theta_0^{(1)} - \theta_0^{(2)}\right\|_2^2\right]. \tag{B.13}$$

What we find is that a one-step application of the kernel on both distributions has a bound with respect to the starting distance. Inducting over this, at the k^{th} step we will have that:

$$\mathrm{W}_2^2(\lambda_1 R_\gamma^k, \lambda_2 R_\gamma^k) \leqslant \mathbb{E}\left[\left\|\theta_k^{(1)} - \theta_k^{(2)}\right\|_2^2\right] \leqslant \left[\underbrace{1 - 2\mu\gamma\left(1 - \gamma\frac{L}{2}\right)}_{:=\hbar}\right]^k \mathrm{W}_2^2(\lambda_1, \lambda_2). \tag{B.14}$$

In words, starting from any distribution, we will get exponentially close to the initial distance as steps go on, if and only if $\gamma \in (0, 2/L)$, which we indeed assumed. Choosing specifically $\lambda_2 = \lambda_1 R_\gamma$ to be one step later,[7] we obtain that:

$$\sum_{k=1}^{N} \mathrm{W}_2^2(\lambda_1 R_\gamma^k, \lambda_2 R_\gamma^k) = \sum_{k=1}^{N} \mathrm{W}_2^2(\lambda_1 R_\gamma^k, \lambda_1 R_\gamma^{k+1}) \leqslant \sum_{k=1}^{N} \hbar^k \mathrm{W}_2^2(\lambda) \tag{B.15}$$

Since we found that the elements are decaying geometrically, the sum on the left-hand side is finite, therefore making the sequence $(\lambda_1 R_\gamma^k)_{k\in\mathbb{N}}$ Cauchy. Obviously, our distance is nice and makes $\mathcal{P}_2(\mathbb{R}^d)$ complete (see Villani (2009, thm. 6.16)), so Cauchyness is characterized by having a limit in $\mathcal{P}_2(\mathbb{R}^d)$, which we call $\pi_\gamma^{\lambda_1} \in \mathcal{P}_2(\mathbb{R}^d)$.

We have more, since from any two starting distributions their distance decays quickly, so we can automatically obtain uniqueness of the limit by contradiction. Indeed, assume there exists another limit $\pi_\gamma^{\lambda_2}$ for $\lambda_2 \neq \lambda_1$, then:

$$\begin{aligned}
\mathrm{W}_2(\pi_\gamma^{\lambda_1}, \pi_\gamma^{\lambda_2}) &\leqslant \mathrm{W}_2(\pi_\gamma^{\lambda_1}, \lambda_1 R_\gamma^k) + W_2(\lambda_1 R_\gamma^k, \pi_\gamma^{\lambda_2}) && \text{tr. ineq.;} && \text{(B.16)}\\
&\leqslant \mathrm{W}_2(\pi_\gamma^{\lambda_1}, \lambda_1 R_\gamma^k) + W_2(\lambda_1 R_\gamma^k, \lambda_2 R_\gamma^k) + W_2(\lambda_2 R_\gamma^k, \pi_\gamma^{\lambda_2}) && \text{tr. ineq.;} && \text{(B.17)}\\
&\overset{k\to\infty}{\to} 0, && && \text{(B.18)}
\end{aligned}$$

since all the three terms go to zero: the outer by assumption, the middle one by the general decaying property we found. By the fact that the Wasserstein is a distance, *a fortiori* we have $\pi_\gamma^{\lambda_1} = \pi_\gamma^{\lambda_2}$ contradicting the hypothesis.

Let us now pass to the quantification of convergence. We aim to apply our master equation eq. (B.14) again, for a particular choice. Taking $\lambda_1 = \delta_\theta, \lambda_2 = \pi_\gamma$, we will have that $\lambda_2 R_\gamma^k = \pi_\gamma R_\gamma^k = \pi_\gamma \in \mathcal{P}_2(\mathbb{R}^d)$ by stationarity, and thus will look at a distance from an arbitrary starting point and the final distribution. Specializing the bound, the claim is proved by definition of Wasserstein distance (def. D.16). Also, the bound is non-trivial since: $\int_{\mathbb{R}^d} \|\theta - \xi\|_2^2 \,\mathrm{d}\pi_\gamma(\xi) \leqslant 2\|\theta\|^2 + 2\int_{\mathbb{R}^d} \|\xi\|_2^2 \,\mathrm{d}\pi_\gamma(\xi) < \infty$.

For the Lipschitz criterion, a little more does the game. With the same particular measures, we evaluate expectations of Lipschitz functions ϕ with L_ϕ constant:

$$\begin{aligned}
\left|R_\gamma^k(\phi(\theta)) - \pi_\gamma\phi\right| &= \left|\mathbb{E}\left[\phi(\theta_{k,\gamma}^{(1)}) - \phi(\theta_{k,\gamma}^{(2)})\right]\right| && && \text{(B.19)}\\
&\leqslant L_\phi \mathbb{E}\left[\left\|\theta_{k,\gamma}^{(1)} - \theta_{k,\gamma}^{(2)}\right\|_2\right] && && \text{(B.20)}\\
&\leqslant L_\phi \sqrt{\mathbb{E}\left[\left\|\theta_{k,\gamma}^{(1)} - \theta_{k,\gamma}^{(2)}\right\|_2^2\right]} && && \text{(B.21)}\\
&\leqslant L_\phi \sqrt{\hbar^k} \sqrt{\int_{\mathbb{R}^d} \|\theta - \xi\| \,\mathrm{d}\pi_\gamma(\xi)} && \text{master equation B.14.} && \text{(B.22)}
\end{aligned}$$

□

[7] We can do this since by A.2-A.3-A.4 it holds that $\lambda_1 R_\gamma \in \mathcal{P}_2(\mathbb{R}^d)$, by an easy triangle inequality and the uniform bound on the gradients basically.

Proposition B.23 (Prop. 17 in (Dieuleveut, Durmus, and Bach 2020)). *Recall the statement for the baby example of proposition 2.6. There, we can further say that for all $\gamma \in (0, 1/r^2)$, where r^2 is defined in appendix C that:*

$$\boldsymbol{R}^{-1} \int_{\mathbb{R}^d} [\boldsymbol{\theta} - \boldsymbol{\theta}^\star][\boldsymbol{\theta} - \boldsymbol{\theta}^\star]^\top \, \mathrm{d}\pi_\gamma(\boldsymbol{\theta}) = \gamma \mathbb{E}\left[\xi_1 \xi_1^\top\right], \tag{B.24}$$

where ξ is the multiplicative error of equation eq. (C.6).

C BABY EXAMPLE DETAILS

In particular, if we take the square loss and a linear model $y = \langle \boldsymbol{x}, \boldsymbol{\theta}^\star \rangle + \eta$, we find $\mathscr{L}(\boldsymbol{x}, y, \boldsymbol{\theta}) = (\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle - y)^2$ which has an explicit generalization error:[8]

$$\mathscr{E}(\boldsymbol{\theta}; \mathbf{L}) = \mathbb{E}_{(\mathbf{x},y)}\left[(\langle \mathbf{x}, \boldsymbol{\theta} \rangle - y)^2\right] = \mathbb{E}_{(\mathbf{x},y)}\left[(\langle \mathbf{x}, \boldsymbol{\theta} - \boldsymbol{\theta}^\star \rangle - \eta)^2\right] = \|\boldsymbol{\Sigma}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\|_2^2 + \sigma^2, \qquad \boldsymbol{\Sigma} := \mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right]. \tag{C.1}$$

Instead, the iterations of SGD will see a gradient that is unbiased:

$$\mathbb{E}_{(x_k,y_k)}\left[\nabla \mathscr{L}\left(\mathbf{x}_k, y_k, \theta_k^{(\gamma)}\right)\right] = \mathbb{E}_{(x_k,\eta_k)}\left[2x_k\left(\left\langle x_k, \theta_k^{(\gamma)} - \boldsymbol{\theta}^\star \right\rangle + \eta_k\right)\right] = 2\boldsymbol{\Sigma}(\theta_k^{(\gamma)} - \boldsymbol{\theta}^\star) = \nabla \mathscr{E}\left(\theta_k^{(\gamma)}; \mathscr{L}\right). \tag{C.2}$$

Moreover, it is interesting to express the gradient in terms of the true gradient by making the right terms appear:

$$\begin{aligned}
\frac{1}{2}\nabla \mathscr{L}\left(\mathbf{x}_k, y_k, \theta_k^{(\gamma)}\right) &= \mathbf{x}_k \mathbf{x}_k^\top \left(\theta_k^{(\gamma)} - \boldsymbol{\theta}^\star\right) - \mathbf{x}_k \eta_k && \text{(C.3)}\\
&= \boldsymbol{\Sigma}\left(\theta_k^{(\gamma)} - \boldsymbol{\theta}^\star\right) + \left[\mathbf{x}_k \mathbf{x}_k^\top - \boldsymbol{\Sigma}\right]\left(\theta_k^{(\gamma)} - \boldsymbol{\theta}^\star\right) - \mathbf{x}_k \eta_k && \text{(C.4)}\\
&= \frac{1}{2}\nabla \mathscr{E}\left(\theta_k^{(\gamma)}; \mathbf{L}\right) + \left[\mathbf{x}_k \mathbf{x}_k^\top - \boldsymbol{\Sigma}\right]\left(\theta_k^{(\gamma)} - \boldsymbol{\theta}^\star\right) - \mathbf{x}_k \eta_k. && \text{(C.5)}
\end{aligned}$$

We can further make sense of the error term by a decomposition into an *additive* and a *multiplicative* term as $\epsilon_k \equiv \epsilon_k\left(\theta_k^{(\gamma)}\right) = \rho_k\left(\theta_k^{(\gamma)}\right) + \xi_k$, where

$$\rho_k \equiv \rho_k(\boldsymbol{\theta}) := \left[\mathbf{x}_k \mathbf{x}_k^\top - \boldsymbol{\Sigma}\right]\left(\theta_k^{(\gamma)} - \boldsymbol{\theta}^\star\right), \qquad \xi_k := -\mathbf{x}_k \eta_k = \left(\mathbf{x}_k^\top \boldsymbol{\theta}^\star - y_k\right)\mathbf{x}_k. \tag{C.6}$$

In particular, the multiplicative noise is independent of the current iterate. Both appear in the proof of proposition 2.10.

ABOUT THE ASSUMPTIONS It is not automatic that the baby example of section 2.II satisfies the assumptions. In particular:

- for $p \geqslant 2$, assumption A.4 will hold;
- assumption A.5 requires almost sure boundedness of observations;
- for A.4 to hold, we might require that there exists $r \geqslant 0$ such that $\mathbb{E}\left[\|\boldsymbol{x}_k\|_2^2 \boldsymbol{x}_k \boldsymbol{x}_k^\top\right] \preceq r^2 \boldsymbol{\Sigma}$, which holds if data is a.s. bounded or has bounded kurtosis, as mentioned in (Dieuleveut, Flammarion, and Bach 2017).

D USEFUL DEFINITIONS AND THEOREMS

We report here some classical definitions for completeness.

Definition D.1 (μ-strong convexity). *A function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex for $\mu > 0$ when:*

$$f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{x}') \leqslant \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{x}') - (1-\alpha)\alpha \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2, \qquad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathscr{X}, \forall \alpha \in [0,1]. \tag{D.2}$$

Definition D.3 (L-smoothness). *A function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if its gradient is L-Lipschitz.*

[8] Let η be square integrable and have variance σ^2.

Lemma D.4 (Descent lemma)**.** *If f is L-smooth then for all $x \in \mathrm{int}(\mathrm{dom}(f))$ and $y \in \mathrm{dom}(f)$ it holds:*

$$f(\boldsymbol{y}) \leqslant f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2. \tag{D.5}$$

Proof. Let $t \in [0, 1]$. Define the interpolating scalar function:

$$\varphi(t) \coloneqq f(\boldsymbol{x} + t\boldsymbol{h}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), t\boldsymbol{h} \rangle, \qquad \boldsymbol{h} \coloneqq \boldsymbol{y} - \boldsymbol{x}. \tag{D.6}$$

In particular, $\varphi(0) = 0$ and we wish to prove $\varphi(1) \leqslant {}^{L}\!/_{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$. Clearly, φ is differentiable since f is. We can bound the derivative as:

$$\varphi'(t) = \langle \boldsymbol{h}, f(\boldsymbol{x} + t\boldsymbol{h}) \rangle - \langle \boldsymbol{h}, \nabla f(\boldsymbol{x}) \rangle = \langle \nabla f(\boldsymbol{x} + t\boldsymbol{h}) - \nabla f(\boldsymbol{x}), \boldsymbol{h} \rangle \leqslant tL\|\boldsymbol{h}\|_2^2. \tag{D.7}$$

We can then apply the fundamental theorem of calculus:

$$f(\boldsymbol{y}) - \varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)\,\mathrm{d}t \leqslant L\|\boldsymbol{h}\|_2^2 \int_0^1 t\,\mathrm{d}t = \frac{L}{2}\|\boldsymbol{h}\|_2^2, \tag{D.8}$$

which proves the claim. □

Combining the first definition and the descent lemma, we find that assuming μ-strong convexity and L-smoothness bounds the function in a variable way as:

$$f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \leqslant f(\boldsymbol{y}) \leqslant f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2, \qquad \forall \boldsymbol{x}, \boldsymbol{y}. \tag{D.9}$$

In words, at each fixed point $\boldsymbol{y}$, we need to bound the function value by any "parabola" parameterized by the auxiliary $\boldsymbol{x}$ vectors. While seemingly irrelevant, this is quite stringent. If we take the special case where $\frac{\mu}{=} L$, then we only have quadratic functions! If we allow for $\mu < L$ strictly, we get a non-trivial class, that is however still very stringent. On a side note, if we reorder terms we can see that our definition is essentially a bound on the convexity gap, which must be quadratic:

$$\frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \leqslant \underbrace{f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle}_{\coloneqq \mathrm{D}_{\mathsf{Bre}}(\boldsymbol{y}|\boldsymbol{x};f)} \leqslant \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2, \tag{D.10}$$

where the middle term is also termed Bregman divergence (hence the notation in underbrace).

Definition D.11 (Locally-polynomially Lipschitz)**.** *A function $f : \mathbb{R}^d \to \mathbb{R}^p$ is locally-polynomially Lipschitz (locally Lipschitz for short) if there exists $\alpha \geqslant 0$ such that:*

$$\left\|f(\boldsymbol{x}) - f(\boldsymbol{x}')\right\|_2 \leqslant \left(1 + \|\boldsymbol{x}\|_2^{\alpha} + \left\|\boldsymbol{x}'\right\|_2^{\alpha}\right) \left\|\boldsymbol{x} + \boldsymbol{x}'\right\|_2, \qquad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d. \tag{D.12}$$

Definition D.13 (C-co-coercivity)**.** *A function $g : \mathbb{R}^d \to \mathbb{R}^p$ is C-co-coercive if for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ we have:*

$$C\left\langle g(\boldsymbol{x}) - g(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}' \right\rangle \geqslant \left\|g(\boldsymbol{x}) - g(\boldsymbol{x}')\right\|_2^2. \tag{D.14}$$

In the main text we assume that the gradient is both L-co-coercive and Lipschitz, by a simple application of the Cauchy-Schwartz inequality we will have that:

$$\left\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}')\right\|_2^2 \leqslant L\left\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}' \right\rangle \leqslant L^2\left\|\boldsymbol{x} - \boldsymbol{x}'\right\|_2^2, \qquad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d. \tag{D.15}$$

Definition D.16 (Order 2-Wasserstein distance)**.** *Let $\lambda, \nu \in \mathcal{P}_2^{(}\mathbb{R}^d)$ be two measures with finite second moment, and $\Pi(\lambda, \nu)$ be the set of couplings on $\mathbb{R}^d \times \mathbb{R}^d$ that have marginals (λ, ν). We define their 2-Wasserstein distance as:*

$$\mathrm{W}_2(\lambda, \nu) \coloneqq \inf_{\xi \in \Pi(\lambda,\nu)} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \; \xi(\,\mathrm{d}\boldsymbol{x},\, \mathrm{d}\boldsymbol{y})\right)^{\frac{1}{2}}. \tag{D.17}$$

Definition D.18 (Markov kernel)**.** *A Markov kernel R on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$ is a map such that:*

(K1) for all $A \in \mathscr{B}(\mathbb{R}^d)$ the mapping $\boldsymbol{\theta} \mapsto R(\boldsymbol{\theta}, A)$ is Borel measurable;

(K2) for all $\boldsymbol{\theta} \in \mathbb{R}^d$ the mapping $A \mapsto R(\boldsymbol{\theta}, A)$ is a probability measure on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$.

In practice, we will use it to abstractly represent how the law of eq. (1.1) changes across iterations. That is, we let R_γ depend on γ the step-size, and starting from an arbitrary $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ (that can be sampled) consider the recursion:

$$R_\gamma^1 := R_\gamma, \qquad R_\gamma^{k+1}(\boldsymbol{\theta}_0, A) = \int_{\mathbb{R}^d} R_\gamma^k(\boldsymbol{\theta}_0, \mathrm{d}\boldsymbol{\theta}) R_\gamma(\boldsymbol{\theta}, A), \qquad \forall \boldsymbol{\theta} \in \mathbb{R}^d, A \in \mathscr{B}(\mathbb{R}^d), k \in \mathbb{N}, \tag{D.19}$$

where the base iterate acts such that almost surely $R_\gamma(\boldsymbol{\theta}_k, A) = \mathbb{P}\left[\boldsymbol{\theta}_{k+1} \in A \mid \boldsymbol{\theta}_k\right]$ for all $k \in \mathbb{N}, \forall A \in \mathscr{B}(\mathbb{R}^d)$. In (informal!) words, the probability that we are in set A if we were at $\boldsymbol{\theta}_k$ before. Such notational construction allows us to define measures and measurable functions as follows. Since by (K1) at fixed A we have a measurable function, we can integrate out for a given $\lambda \in \mathcal{P}(\mathbb{R}^d)$ as:

$$\lambda R_\gamma^k(\cdot) := \int_{\mathbb{R}^d} \lambda(\mathrm{d}\boldsymbol{\theta}) R_\gamma^k(\boldsymbol{\theta}, \cdot) : \mathscr{B}(\mathbb{R}^d) \to \mathcal{P}(\mathbb{R}^d) \tag{D.20}$$

$$A \mapsto \lambda R_\gamma^k(A) = \mathbb{P}\left[\theta_k^{(\gamma)} \in A \mid \theta_0 \sim \lambda\right]. \tag{D.21}$$

Namely, if we chain from the left we obtain the probability of starting from λ and ending in A at the k^{th} step. If instead we chain from the right, we will obtain a measurable function representing the expectation at the k^{th} step with respect to the iterated distribution. Mathematically, for $\phi \in \mathrm{F}_+(\mathbb{R}^d, \mathbb{R})$ the space of positive measurable functions we have that:

$$R_\gamma^k \phi(\cdot) := \int_{\mathbb{R}^d} \phi(\boldsymbol{\theta}) R_\gamma^k(\cdot, \mathrm{d}\boldsymbol{\theta}) : \mathbb{R}^d \to \mathrm{F}_+(\mathbb{R}^d, \mathbb{R}) \tag{D.22}$$

$$\boldsymbol{\theta}_0 \mapsto R_\gamma^k \phi(\boldsymbol{\theta}_0) = \mathbb{E}\left[\phi\left(\theta_k^{(\gamma)}\right) \mid \boldsymbol{\theta}_0\right]. \tag{D.23}$$

We therefore obtain probabilities integrating from the left and expectations integrating from the right. Note that by the notations it follows also that for all $A \in \mathscr{B}(\mathbb{R}^d)$:

$$\lambda\left(R_\gamma^k h\right)(\cdot) = \int_{\mathbb{R}^d} \lambda(\mathrm{d}\boldsymbol{\theta})(R_\gamma^k h)(\boldsymbol{\theta}, \cdot) \tag{D.24}$$

$$= \int_{\mathbb{R}^d} \lambda(\mathrm{d}\boldsymbol{\theta}) \int_{\mathbb{R}^d} h(\mathrm{d}\boldsymbol{\xi}) R_\gamma^k(\boldsymbol{\theta}, \mathrm{d}\boldsymbol{\xi}) \tag{D.25}$$

$$= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \lambda(\mathrm{d}\boldsymbol{\theta}) R_\gamma^k(\boldsymbol{\theta}, \mathrm{d}\boldsymbol{\xi}) h(\boldsymbol{\xi}) \tag{D.26}$$

$$= \int_{\mathbb{R}^d} h(\boldsymbol{\xi}) \int_{\mathbb{R}^d} \lambda(\mathrm{d}\boldsymbol{\theta}) R_\gamma^k(\boldsymbol{\theta}, \mathrm{d}\boldsymbol{\xi}) \tag{D.27}$$

$$= \int_{\mathbb{R}^d} h(\boldsymbol{\theta})(\lambda R_\gamma^k)(\cdot, \mathrm{d}\boldsymbol{\theta}) \tag{D.28}$$

$$= (\lambda R_\gamma^k)(h)(\cdot). \tag{D.29}$$

Proposition D.30 (Theorem 6.14 in (Le Gall 2016))**.** *Let h, g be continuous functions on a euclidean domain $E \subset \mathbb{R}^d$ that tend to zero at infinity.*[9] *The following are equivalent:*

1. *$h \in \mathrm{D}(\mathcal{A})$ and $\mathcal{A}h = g$;*
2. *for all $\boldsymbol{x} \in E$ the process:*

$$h(\theta_t^{\boldsymbol{\theta}_0}) - \int_0^t g(\theta_s^{\boldsymbol{\theta}_0})\,\mathrm{d}s, \tag{D.31}$$

is a martingale with respect to the (canonical) filtration $(\mathcal{F}_t)_{t\in[0,\infty]}$. Here, by $\theta_t^{\boldsymbol{\theta}_0}$ we mean the Markov chain $(\theta_t)_{t\in\mathbb{R}_+}$ that started almost surely at $\boldsymbol{\theta}_0$.

[9] This is a specific definition: for all $\epsilon > 0$ there exists a compact set K such that $|f(\boldsymbol{\theta})| < \epsilon$ for all $\boldsymbol{\theta} \in E \backslash K$.