Taming neural networks with TUSLA: Non-convex learning via adaptive stochastic gradient Langevin algorithms *

Attila Lovas[1], Iosif Lytras[2], Miklós Rásonyi[1], and Sotirios Sabanis[2, 3]

[1] Alfréd Rényi Institute of Mathematics, 1053 Budapest, Reáltanoda utca 13–15, Hungary
[2] School of Mathematics, The University of Edinburgh, UK.
[3] The Alan Turing Institute, UK.

June 26, 2020

Abstract

Artificial neural networks (ANNs) are typically highly nonlinear systems which are finely tuned via the optimization of their associated, non-convex loss functions. Typically, the gradient of any such loss function fails to be dissipative making the use of widely-accepted (stochastic) gradient descent methods problematic. We offer a new learning algorithm based on an appropriately constructed variant of the popular stochastic gradient Langevin dynamics (SGLD), which is called tamed unadjusted stochastic Langevin algorithm (TUSLA). We also provide a nonasymptotic analysis of the new algorithm's convergence properties in the context of non-convex learning problems with the use of ANNs. Thus, we provide finite-time guarantees for TUSLA to find approximate minimizers of both empirical and population risks. The roots of the TUSLA algorithm are based on the taming technology for diffusion processes with superlinear coefficients as developed in Sabanis (2013, 2016) and for MCMC algorithms in Brosse et al. (2019). Numerical experiments are presented which confirm the theoretical findings and illustrate the need for the use of the new algorithm in comparison to vanilla SGLD within the framework of ANNs.

1 Introduction

A new generation of stochastic gradient decent algorithms, namely stochastic gradient Langevin dynamics (SGLD), can be efficient in finding global minimizers of possibly complicated, high-dimensional landscapes under suitable regularity assumptions for the gradient, see Raginsky et al. (2017), Welling and Teh (2011) and references therein. However, in the specific case of tuning ANNs, or simply neural networks henceforth, problems arise already at the theoretical level. As discussed in Section 4 below in some detail, the functionals to be minimized fail any form of dissipativity which should be a *sine qua non* for any stable gradient algorithms. Adding a quadratic regularization term cannot always remedy this, in which case one needs to replace it with a higher order penalty term. However, the addition of such a term leads to the violation of the global Lipschitz continuity for the regularized gradient, which in turn renders the use of gradient descent methods problematic as it can be seen in Figure 2. This issue has been highlighted in the case of Euler discretizations (of which SGLD is an example) in Hutzenthaler et al. (2011), where it is proven that the difference of the exact solution of the corresponding stochastic differential equation (SDE) and of the numerical approximation at even a finite time point diverges to infinity in the strong mean square sense.

A natural way to address the above issue is to combine higher order regularization with taming techniques to improve the stability of any resulting algorithm. In particular, the use of taming techniques in the construction of stable numerical approximations for nonlinear SDEs has gained substantial attention in recent years and was introduced by Hutzenthaler et al. (2012) and, independently,

* All the authors were supported by The Alan Turing Institute, London under the EPSRC grant EP/N510129/1. A. L. and M. R. thank for the "Lendület" grant LP 2015-6 of the Hungarian Academy of Sciences.

by Sabanis (2013, 2016). The latter taming approach was used in the creation of a new generation of Markov chain Monte Carlo (MCMC) algorithms, see Brosse et al. (2019), Sabanis and Zhang (2019), which are designed to sample from distributions such that the gradient of their log density is only locally Lipschitz continuous and is allowed to grow superlinearly at infinity.

It is essential here to recall the importance of Langevin based algorithms. Their nonasymptotic convergence analysis has been highlighted in recent years by numerous articles in the literature. For the case of deterministic gradients one could consult Dalalyan (2017), Durmus and Moulines (2017, 2019), Cheng et al. (2018), Sabanis and Zhang (2019) and references therein, whereas for stochastic gradients of convex potentials details can be found in Brosse et al. (2018), Dalalyan and Karagulyan (2019) and in Barkhagen et al. (2018) which goes beyond the case of iid data. Further, due to the newly obtained results in the study of contraction rates for Langevin dynamics, see Eberle et al. (2019b,a), the case of nonconvex potentials within the framework of stochastic gradients was studied in Raginsky et al. (2017), Xu et al. (2018) and, in particular, substantial progress has been made in Chau et al. (2019) by obtaining the best known convergence rates even in the presence of dependent data streams. The latter article has inspired the development of the SGLD theory under local conditions, see Zhang et al. (2019), which provides theoretical convergence guarantees for a wide class of applications, including scalable posterior sampling for Bayesian inference and nonconvex optimization arising in variational inference problems.

Despite all this very significant progress, the use of SGLD algorithms for the fine tuning of neural networks remained only at a heuristic level without any theoretical guarantees for the discovery of approximate minimizers of empirical and population risks. To the best of the authors' knowledge, the current article is the first work to address this shortcoming in the theory of Langevin algorithms by presenting a novel new algorithm, which is called tamed unadjusted stochastic Langevin algorithm (TUSLA), along with a nonasymptotic analysis of its convergence properties.

We conclude this section by introducing some notation. Let $(\Omega, \mathcal{F}, P)$ be a probability space. We denote by $\mathbb{E}[X]$ the expectation of a random variable X. For $1 \leq p < \infty$, L^p is used to denote the usual space of p-integrable real-valued random variables. Fix an integer $d \geq 1$. For an $\mathbb{R}^d$-valued random variable X, its law on $\mathcal{B}(\mathbb{R}^d)$, i.e. the Borel sigma-algebra of $\mathbb{R}^d$, is denoted by $\mathcal{L}(X)$. Scalar product is denoted by $\langle \cdot, \cdot \rangle$, with $|\cdot|$ standing for the corresponding norm (where the dimension of the space may vary depending on the context). For $\mu \in \mathcal{P}(\mathbb{R}^d)$ and for a non-negative measurable $f : \mathbb{R}^d \to \mathbb{R}$, the notation $\mu(f) := \int_{\mathbb{R}^d} f(\theta)\mu(\mathrm{d}\theta)$ is used. For any integer $q \geq 1$, let $\mathcal{P}(\mathbb{R}^q)$ denote the set of probability measures on $\mathcal{B}(\mathbb{R}^q)$. For $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, let $\mathcal{C}(\mu, \nu)$ denote the set of probability measures ζ on $\mathcal{B}(\mathbb{R}^{2d})$ such that its respective marginals are μ, ν. For two probability measures μ and ν, the Wasserstein distance of order $p \geq 1$ is defined as

$$W_p(\mu, \nu) := \inf_{\zeta \in \mathcal{C}(\mu,\nu)} \left(\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |\theta - \theta'|^p \zeta(\mathrm{d}\theta \mathrm{d}\theta') \right)^{1/p}, \ \mu, \nu \in \mathcal{P}(\mathbb{R}^d). \tag{1}$$

2 Main results and assumptions

We consider initially the setting which is required for the precise formulation of the newly proposed algorithm. To this end, let us denote by $(\mathcal{G}_n)_{n\in\mathbb{N}}$ a given filtration representing the flow of past information. Moreover, let $(X_n)_{n\in\mathbb{N}}$ be an $\mathbb{R}^m$-valued, $(\mathcal{G}_n)$-adapted process and $(\xi_n)_{n\in\mathbb{N}}$ be an $\mathbb{R}^d$-valued Gaussian process. It is assumed throughout the paper that the random variable θ_0 (initial condition), $\mathcal{G}_\infty$ and $(\xi_n)_{n\in\mathbb{N}}$ are independent. Let also $G : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ be a continuously differentiable function. The required assumptions are as follows.

2.1 Assumptions and key observations

Although the assumptions below are presented in a formal way for the general case of locally Lipschitz continuous gradients, the connection with neural networks is given explicitly in Section 4. In particular, the function G below can be seen as the stochastic gradient described in equation (19).

Assumption 1. *There exist positive constants L_1, ρ and $q \geq 1$ such that*

$$\left|G(\theta, x) - G\left(\theta', x\right)\right| \leq L_1(1+|x|)^{\rho}(1+|\theta|+|\theta'|)^{q-1}\left|\theta - \theta'\right|, \text{ for all } x \in \mathbb{R}^m \text{ and } \theta, \theta' \in \mathbb{R}^d.$$

Definition 2.1. *Let $\eta \in (0,1)$ be a reguralization parameter and r be a constant such that $r \geq \frac{q}{2}+1$. Then, the stochastic gradient with the necessary regularised term is given by*

$$H(\theta, x) := G(\theta, x) + \eta\theta|\theta|^{2r}$$

for all $x \in \mathbb{R}^m$ and $\theta \in \mathbb{R}^d$. Moreover, $g(\theta) := \mathbb{E}\left[G\left(\theta, X_0\right)\right]$ and $h(\theta) := \mathbb{E}\left[H(\theta, X_0)\right]$ for every $\theta \in \mathbb{R}^d$.

Remark 2.2. *As an example, H can be seen as the gradient of a function of the form*

$$U(\theta, x) := F(\theta, x) + \frac{\eta}{2(r+1)}|\theta|^{2(r+1)}, \text{ where } G(\theta, x) := \nabla_\theta F(\theta, x), \text{ for all } \theta \in \mathbb{R}^d, x \in \mathbb{R}^m.$$

Assumption 2. *The process $(X_n)_{n\geq 1}$ is a sequence of i.i.d. random variables with $\mathbb{E}\left|X_0\right|^{16\rho(2r+1)} < \infty$, where ρ is given in Assumption 1 and r in Definition 2.1. In addition, the initial condition is such that $\mathbb{E}\left|\theta_0\right|^{16(2r+1)} < \infty$.*

Remark 2.3. *By taking a closer look at Assumption 1, one observes that the growth of G can be controlled, i.e. for every $\theta \in \mathbb{R}^d$ and $x \in \mathbb{R}^m$*

$$|G(\theta, x)| \leq K(x)(1+|\theta|^q), \tag{2}$$

where $K(x) = 2^q\left(L_1(1+|x|)^{\rho} + |G(0,x)|\right)$.

Remark 2.4. *In view of Assumptions 1 and 2, one obtains that*

$$\langle\theta, h(\theta)\rangle = \langle\theta, \mathbb{E}\, G(\theta, X_0)\rangle + \langle\theta, \eta\theta|\theta|^{2r}\rangle \geq \eta|\theta|^{2r+2} - \mathbb{E}[K(X_0)]|\theta|(1+|\theta|^q).$$

Furthermore, for $A = \mathbb{E}[K(X_0)]$ and $B = \left(3\,\mathbb{E}[K(X_0)]\right)^{q+2}\eta^{-q-1}$, it holds that

$$\langle\theta, h(\theta)\rangle \geq A|\theta|^2 - B. \tag{3}$$

Remark 2.5. *Assumption 1 yields that $\langle\theta, \mathbb{E}\left[G(\theta, X_0)\right]\rangle \geq -\,\mathbb{E}\left[K(X_0)\right]\left(|\theta| + |\theta|^{q+1}\right)$.*

Proposition 2.6. *Let Assumptions 1 and 2 hold. Then, for every $\theta, \theta' \in \mathbb{R}^d$,*

$$\langle\theta - \theta', h(\theta) - h(\theta')\rangle \geq -a|\theta - \theta'|^2,$$

where $a = \sqrt{d}L(1+2|R|)^{q-1}$ and R is given explicitly in the proof.

The following proposition states that the stochastic gradient is not globally Lipschitz continuous in θ, hence a new approach is required for learning schemes which rely on the analysis of Langevin dynamics with gradients satisfying weaker smoothness conditions. Crucially though, the local Lipschitz continuity property remains true and, moreover, the associated local Lipschitz constant is controlled by powers of the state variables which allow us to used an approach based on taming techniques.

Proposition 2.7. *Let Assumptions 1 and 2 hold. Then, in view of Definition 2.1 one obtains that*

$$|H(\theta, x) - H(\theta', x)| \leq L(1+|x|)^{\rho}(1+|\theta|+|\theta'|)^{l}|\theta - \theta'|, \text{ for all } x \in \mathbb{R}^m, \text{ and } \theta, \theta' \in \mathbb{R}^d$$

where $L = L_1 + 8r\eta$ and $l = 2r+1$.

2.2 The new algorithm and main results

We introduce a new iterative scheme, which is a hybrid of the stochastic gradient Langrevin dynamics (SGLD) algorithm and of the tamed unadjusted Langevin algorithm and uses 'taming', see Sabanis (2013, 2016), Brosse et al. (2019) and references therein, for asserting control on the superlinearly growing gradient. This new algorithm is called TUSLA, tamed unadjusted stochastic Langevin algorithm, and is given by

$$\theta_{n+1}^{\lambda} := \theta_n^{\lambda} - \lambda H_\lambda(\theta_n^{\lambda}, X_{n+1}) + \sqrt{2\lambda\beta^{-1}}\xi_{n+1},\ n \in \mathbb{N}, \tag{4}$$

where $\theta_0^{\lambda} := \theta_0$ and

$$H_\lambda(\theta, x) := \frac{H(\theta, x)}{1 + \sqrt{\lambda}|\theta|^{2r}}, \qquad \text{for every } \theta \in \mathbb{R}^d, x \in \mathbb{R}^m, \tag{5}$$

where $\{\xi_n\}_{n\geq 1}$ is a sequence of independent standard d-dimensional Gaussian random variables. The new algorithm addresses known stability issues of SGLD algorithms, see also Figure 2, and can be seen as an SGLD algorithm with adaptive step size. This is due to the fact that, at each iteration, the stochastic gradient H is multiplied with a step size which is controlled by the $2r$-th power of the (vector) norm of the parameter, i.e. by $\lambda\left(1 + \sqrt{\lambda}|\theta|^{2r}\right)^{-1}$.

Henceforth, λ is assumed to be controlled by

$$\lambda_{max} = \min\{1, \frac{1}{4\eta^2\left(8(p+1)\left(\lceil\frac{p}{2}\rceil\right)^2\right)^2}, \frac{1}{4\eta^2}\} \tag{6}$$

where p depends on which $2p$-th moment of θ_n we need to estimate.

Remark 2.8. *Observe that, due to Remark 2.3 and (5),*

$$\mathbb{E}[\sqrt{\lambda}|H_\lambda(\theta_n^{\lambda}, X_{n+1})|\theta_n^{\lambda}] \leq \sqrt{\lambda}\frac{\mathbb{E}\left[K(X_0)\right](1 + |\theta_n^{\lambda}|^q) + \eta|\theta_n^{\lambda}|^{2r+1}}{1 + \sqrt{\lambda}|\theta|^{2r}} \leq \mathbb{E}\left[K(X_0)\right] + \eta|\theta_n^{\lambda}|. \tag{7}$$

Moreover,

$$\mathbb{E}[\lambda|H_\lambda(\theta_n^{\lambda}, X_{n+1})|^2|\theta_n^{\lambda}] \leq 4\,\mathbb{E}[K^2(X_0)] + 2\eta^2|\theta_n^{\lambda}|^2. \tag{8}$$

It is well-known that, under mild conditions, which in this case are satisfied due to Assumptions 1–2 and, in particular, due to (3), the so-called (overdamped) Langevin SDE which is given by

$$\mathrm{d}Z_t = -h\left(Z_t\right)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B_t, \quad t > 0 \tag{9}$$

with a (possibly random) initial condition θ_0 and with B_t denoting a d-dimensional Brownian motion, admits a unique invariant measure π_β.

The two main results are given below with regards to the convergence of TUSLA (4) to π_β in metrics W_1 and W_2 as defined in (1).

Theorem 2.9. *Let Assumptions 1 and 2 hold. Then, there exist positive constants C_1, C_2, $\hat{c}$, $\dot{c}$ and z_1 such that, for every $0 < \lambda \leq \lambda_{\max}$,*

$$W_1\left(\mathcal{L}\left(\theta_n^{\lambda}\right), \pi_\beta\right) \leq \sqrt{\lambda}(z_1 + \sqrt{e^{3a}(C_1 + C_2)}) + \hat{c}e^{-\dot{c}n}\left[1 + \mathbb{E}\left[V_2\left(\theta_0\right)\right] + \int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(d\theta)\right],$$

where V_2 is defined in (11). The constants are given explicitly in the proof.

Corollary 2.10. *Let Assumptions 1 and 2 hold. Then, there exist positive constants C_1, C_2 and z_2 such that, for every $0 < \lambda \leq \lambda_{max}$,*

$$W_2\left(\mathcal{L}\left(\theta_n^{\lambda}\right), \pi_\beta\right) \leq \sqrt{e^{3a}(C_1 + C_2)}\sqrt{\lambda} + z_2\lambda^{\frac{1}{4}} + \sqrt{2\hat{c}e^{-\dot{c}n}\left[1 + \mathbb{E}\left[V_2\left(\theta_0\right)\right] + \int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(d\theta)\right]},$$

where V_2 is defined in (11). The constants are given explicitly in the proof.

If we further assume the setting of Remark 2.2, where $h := \nabla u$ with $u(\theta) = \mathbb{E}[U(\theta, X_0)] \geq 0$, then the following non-convex optimization problem can be formulated

$$\text{minimize} \quad u(\theta) := \mathbb{E}[U(\theta, X_0)],$$

where $\theta \in \mathbb{R}^d$ and X_0 is a random element with some unknown probability law. One then needs to estimate a $\hat{\theta}$, more precise its law, such that the expected excess risk $\mathbb{E}[u(\hat{\theta})] - \inf_{\theta \in \mathbb{R}^d} u(\theta)$ is minimized. This optimization problem can thus be decomposed into subproblems, see Raginsky et al. (2017), one of which is a problem of sampling from the target distribution $\pi_\beta(\theta) \propto \exp(-\beta u(\theta))$ with $\beta > 0$. The results in Theorem 2.9 and Corollary 2.10 provide the estimates for this sampling problem. Moreover, at an intuitive level, one understands that the two problems, namely sampling and optimization, are linked in this case since π_β concentrates around the minimizers of u when β takes sufficiently large values, see Hwang (1980) for more details. In fact, one observes that if θ_n^λ is used in place of $\hat{\theta}$, then expected excess risk can be estimated as follows

$$\mathbb{E}\left[u\left(\theta_n^\lambda\right)\right] - u_\star = \underbrace{\mathbb{E}\left[u\left(\theta_n^\lambda\right)\right] - \mathbb{E}\left[u\left(\theta_\infty\right)\right]}_{\mathcal{T}_1} + \underbrace{\mathbb{E}\left[u\left(\theta_\infty\right)\right] - u_\star}_{\mathcal{T}_2} \tag{10}$$

where $u_\star := \inf_{\theta \in \mathbb{R}^d} u(\theta)$. Moreover, the estimates for $\mathcal{T}_1$ rely on the W_2 estimates of Corollary 2.10 and the estimates for $\mathcal{T}_2$ on the properties of the corresponding Gibbs algorithm, see (Raginsky et al., 2017, Section 3.5).

Theorem 2.11. *Let Assumptions 1 and 2 hold. Then,*

$$\mathbb{E}\left[u\left(\theta_n^\lambda\right)\right] - u_\star \leq \left(\frac{a_1}{l+1}\sqrt{\mathbb{E}\,|\theta_0|^{2l} + C\text{`}_l} + \frac{a_1}{l+1}\sqrt{\sigma_{2l}} + r_2\right) W_2\left(\mathcal{L}\left(\theta_n^\lambda\right), \pi_\beta\right) + \left(\frac{eM}{A}\left(\frac{b\beta}{d} + 1\right)\right) - \frac{1}{\beta}\log\left(1 - \frac{d}{M\beta R_0^2}\right),$$

where $a_1 = 2^l(\mathbb{E}[K(X_0)] + \eta)$, $r_2 = 2\,\mathbb{E}[K(X_0)]$, σ_{2l} *is the* $2l$*-moment of* π_β,

$$R_0 = \inf\{y \geq \sqrt{B/A} : \quad y^2(1+4y)^l > \frac{d+1}{\beta L\,\mathbb{E}(1+|X_0|)^\rho}\}$$

$M = L\mathbb{E}(1+|X_0|)^\rho(1+4R_0)^l$ *and* $W_2\left(\mathcal{L}\left(\theta_n^\lambda\right), \pi_\beta\right)$ *is given in Corollary 2.10.*

3 Preliminary estimates

At this point the necessary moments estimates are presented, which guarantee the stability of the new algorithm, along with the necessary (for the approach taken in the proof of the main results) auxiliary processes.

Lemma 3.1. *Let Assumption 1 and 2 hold. For all* $n \in \mathbb{N}$, $p \in [1,\, 8(2r+1)]$ *and* $0 < \lambda < \lambda_{max}$,

$$\mathbb{E}\,|\theta_{n+1}^\lambda|^{2p} \leq (1 - \lambda\eta^2)^n\,\mathbb{E}\,|\theta_0|^{2p} + C\text{`}_p \qquad \textit{and, thus,} \qquad \sup_n \mathbb{E}\,|\theta_n^\lambda|^{2p} < \mathbb{E}\,|\theta_0|^{2p} + C\text{`}_p,$$

where $C\text{`}_p$ *is given explicitly in the proof.*

Before proceeding with the detailed calculations regarding the convergence properties of TUSLA, a suitable family of Lyapunov functions is introduced. For each $m \geq 1$, define the Lyapunov function V_m by

$$V_m(\theta) := \left(1 + |\theta|^2\right)^{m/2}, \quad \theta \in \mathbb{R}^d, \tag{11}$$

and similarly $v_m(x) = (1+x^2)^{\frac{m}{2}}$ for any real $x \geq 0$. Both functions are continuously differentiable and $\lim_{|\theta|\to\infty} \nabla V_m(\theta)/V_m(\theta) = 0$.

Definition 3.2. *We define the continuous-time interpolation of TUSLA, see (4), as*

$$\mathrm{d}\bar{\theta}_t^\lambda = -\lambda H_\lambda\left(\bar{\theta}_{\lfloor t\rfloor}^\lambda, X_{\lfloor t\rfloor}\right)\mathrm{d}t + \sqrt{2\lambda\beta^{-1}}\mathrm{d}\tilde{B}_t^\lambda \tag{12}$$

with initial condition $\bar{\theta}_0^\lambda = \theta_0^\lambda$.

Remark 3.3. *Moreover, due to the homogeneous nature of the coefficients of the continuous-time interpolation of the TUSLA algorithm, the law of the interpolated process (12) coincides with the law of TUSLA (4) at grid points, i.e.* $\mathcal{L}\left(\bar{\theta}_n^\lambda\right) = \mathcal{L}\left(\theta_n^\lambda\right), \quad \forall n \in \mathbb{N}$. *Combining this with the bounds obtained in Lemmas 3.1, one deduces that under the same assumptions,*

$$\sup_{t\geq 0}\mathbb{E}\left|\bar{\theta}_{\lfloor t\rfloor}^\lambda\right|^{2p} \leq \mathbb{E}\left|\theta_0\right|^{2p} + C\text{‘}_p. \tag{13}$$

Furthermore consider a continuous-time process $\zeta_t^{s,v,\lambda}, t \geq s$ which is the solution to the SDE

$$\mathrm{d}\zeta_t^{s,v,\lambda} = -\lambda h\left(\zeta_t^{s,v,\lambda}\right)\mathrm{d}t + \sqrt{2\lambda\beta^{-1}}\mathrm{d}\tilde{B}_t^\lambda \tag{14}$$

with initial condition $\zeta_s^{s,v,\lambda} := v, v \in \mathbb{R}^d$. Let $T := \lfloor 1/\lambda\rfloor$.

Definition 3.4. *Fix* $n \in \mathbb{N}$ *and define* $\bar{\zeta}_t^{\lambda,n} := \zeta_t^{nT,\bar{\theta}_{nT}^\lambda,\lambda}$ *where* $\zeta_t^{nT,\bar{\theta}_{nT}^\lambda,\lambda}$ *is defined in (14).*

Henceforth, any constant denoted by $C\text{‘}_p$, for $p \geq 1$, is given explicitly in the proof of Lemma (3.1).

Lemma 3.5. *Let Assumptions 1 and 2 hold. Then, for* $0 < \lambda < \lambda_{\max}$

$$\mathbb{E}\left[V_4\left(\bar{\theta}_{nT}^\lambda\right)\right] \leq 2(1-\lambda\eta^2)^{nT}\mathbb{E}|\theta_0|^4 + 2 + 2C\text{‘}_2.$$

Lemma 3.6. *Let Assumption 2 holds. Then, for any* $p \geq 2, \theta \in \mathbb{R}^d$,

$$\Delta V_p/\beta - \langle h(\theta), \nabla V_p(\theta)\rangle \leq -\bar{c}(p)V_p(\theta) + \tilde{c}(p),$$

where $\bar{M}_p = \sqrt{1/3 + 4B/(3A) + 4d/(3A\beta) + 4(p-2)/(3A\beta)}$, $\tilde{c}(p) = (3/4)Ap v_p\left(\bar{M}_p\right)$, $\bar{c}(p) = Ap/4$ *and* A, B *are given explicitly in the proof.*

Lemma 3.7. *Let Assumptions 1 and 2 hold. Then,*

$$\mathbb{E}\left[V_2\left(\bar{\zeta}_t^{\lambda,n}\right)\right] \leq \mathbb{E}\left[V_2\left(\theta_0\right)\right] + \frac{\tilde{c}(2)}{\bar{c}(2)} + 2\left(C_X\eta^{-1} + 2M_0^2(2+\eta) + 2d(\eta\beta)^{-1}\sqrt{\lambda_{max}}\right) + 1, \text{ and}$$

$$\mathbb{E}\left[V_4\left(\bar{\zeta}_t^{\lambda,n}\right)\right] \leq 2\mathbb{E}|\theta_0|^4 + 2 + 2C\text{‘}_2 + \frac{\tilde{c}(4)}{\bar{c}(4)}.$$

3.1 Proof of main results

We mainly present the proof of Theorem 2.9. The goal is to establish a non-asymptotic bound for $W_1(\mathcal{L}(\theta_n^\lambda), \pi_\beta)$, which can be split as follows: $W_1(\mathcal{L}(\theta_n^\lambda), \pi_\beta) \leq W_1(\mathcal{L}(\bar{\theta}_n^\lambda), \mathcal{L}(Z_n^\lambda)) + W_1(\mathcal{L}(Z_n^\lambda), \pi_\beta)$. To achieve this, a functional which is associated with the contraction results in Eberle et al. (2019a) and is crucial for obtaining convergence rate estimates in W_1 and W_2. Let $\mathcal{P}_{V_2}$ denote the subset of $\mathcal{P}(\mathbb{R}^d)$ such that every $\mu \in \mathcal{P}_{V_2}$ satisfies $\int_{\mathbb{R}^d} V_2(\theta)\mu(d\theta) < \infty$. The functional is given by

$$w_{1,2}(\mu,\nu) := \inf_{\zeta\in\mathcal{C}(\mu,\nu)}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\left[1 \wedge |\theta-\theta|'\right]\left[\left(1 + V_2(\theta) + V_2\left(\theta'\right)\right)\zeta\left(\mathrm{d}\theta\mathrm{d}\theta'\right)\right. \tag{15}$$

where $\mathcal{C}(\mu,\nu)$ is defined immediately before (1). We can now proceed with the statement of the contraction property of the Langevin SDE (9) in $w_{1,2}$, which yields the desired result for $W_1(\mathcal{L}(Z_n^\lambda), \pi_\beta)$.

Proposition 3.8. *Let $Z'_t, t \in \mathbb{R}_+$ be the solution of the Langevin SDE (9) with initial condition $Z'_0 = \theta_0$ which is independent of $\mathcal{G}_\infty$ and $|\theta_0| \in L^2$. Then,*

$$w_{1,2}\left(\mathcal{L}\left(Z_t\right), \mathcal{L}\left(Z'_t\right)\right) \leq \hat{c}e^{-\dot{c}t} w_{1,2}\left(\mathcal{L}\left(\theta_0\right), \mathcal{L}\left(\theta'_0\right)\right)$$

where $w_{1,2}$ is defined in (15).

The following two Lemmas combined establish the required $W_1(\mathcal{L}(\bar{\theta}^\lambda_n), \mathcal{L}(Z^\lambda_n))$ estimate.

Lemma 3.9. *Let Assumptions 1 and 2 hold. For $0 < \lambda < \lambda_{max}$ and $t \in [nT, (n+1)T]$,*

$$W_2\left(\mathcal{L}\left(\bar{\theta}^\lambda_t\right), \mathcal{L}\left(\bar{\zeta}^{\lambda,n}_t\right)\right) \leq \sqrt{\lambda}\sqrt{e^{3a}(C_1 + C_2)}$$

where C_1, C_2 are given explicitly in the proof.

Lemma 3.10. *Let Assumptions 1 and 2 hold. For $0 < \lambda \leq \lambda_{max}$ and $t \in [nT, (n+1)T]$,*

$$W_1\left(\mathcal{L}\left(\bar{\zeta}^{\lambda,n}_t\right), \mathcal{L}\left(Z^\lambda_t\right)\right) \leq \sqrt{\lambda} z_1$$

where z_1 is given explicitly in the proof.

Thus, in view of the above results, and the facts that $W_1(\mu,\nu) \leq w_{1,2}(\mu,\nu)$ and $\mathcal{L}(\bar{\theta}^\lambda_n) = \mathcal{L}(\theta^\lambda_n)$, for each $n \in \mathbb{N}$, one obtains the results of Theorem 2.9. The proof of Corollary 2.10 follows the same lines by noticing $W_2 \leq \sqrt{2w_{1,2}}$. Full details of all the aforementioned derivations can be found in the Appendix, Section A.3.

Finally, the excess risk as described in (10) is controlled thanks to the following two Lemmas.

Lemma 3.1. *Under the assumptions of the main theorems, there holds*

$$\mathcal{T}_1 := \mathbb{E}[u(\theta^\lambda_n)] - \mathbb{E}[u(\theta_\infty)] \leq \left(\frac{a_1}{l+1}\sqrt{\mathbb{E}\,|\theta_0|^{2l} + C'_l} + \frac{a_1}{l+1}\sqrt{\sigma_{2l}} + r_2\right) W_2\left(\mathcal{L}\left(\theta^\lambda_n\right), \pi_\beta\right)$$

where $a_1 = 2^l(\mathbb{E}\,K(X_0) + \eta)$ and $r_2 = 2\,\mathbb{E}\,K(X_0)$.

Lemma 3.2. *Let $R_0 = \inf\{y \geq \sqrt{B/A} : \quad y^2(1+4y)^l > \frac{d+1}{\beta L\,\mathbb{E}(1+|X_0|)^\rho}\}$ and Assumptions 1 and 2 hold. Then,*

$$\mathcal{T}_2 := \mathbb{E}[u(\theta_\infty)] - u_* \leq \frac{d}{2\beta}\log\left(\frac{eM}{A}\left(\frac{B\beta}{d} + 1\right)\right) - \frac{1}{\beta}\log\left(1 - \frac{d}{M\beta R_0^2}\right).$$

Proof of Theorem 2.11. Due to (10), Lemma 3.1 and Lemma 3.2, the desired result is obtained. □

4 Multilayer neural networks

Some further notation is introduced in this section. The set $\mathbb{N}_+ := \mathbb{N} \setminus \{1\}$ and $\mathrm{id}_{\mathbb{R}^k}$ denotes the identity operator of $\mathbb{R}^k$, $k \in \mathbb{N}$. For $k, l \in \mathbb{N}$, $\mathrm{Lin}\left(\mathbb{R}^k, \mathbb{R}^l\right)$ stands for the vector space of $\mathbb{R}^k \to \mathbb{R}^l$ linear operators. In particular, $(\mathbb{R}^k)^*$ denotes $\mathrm{Lin}\left(\mathbb{R}^k, \mathbb{R}\right)$, that is the dual space of $\mathbb{R}^k$. In our setting, linear functionals and vectors are identified through the inner product. Moreover, for a fixed $v \in \mathbb{R}^k$, we define $M_v \in \mathrm{Lin}\left(\mathbb{R}^k, \mathbb{R}^k\right)$ the element-wise multiplication by v, i.e. $[M_v z]_l = v_l z_l$, $l = 1, \ldots, k$. Furthermore, for an arbitrary $W \in \mathrm{Lin}\left(\mathbb{R}^k, \mathbb{R}^l\right)$, $\|W\|$ stands for the corresponding operator norm, that is $\|W\| = \sup_{|z|=1} |Wz|$. Also, for an arbitrary $W \in \mathrm{Lin}\left(\mathbb{R}^k, \mathbb{R}^l\right)$, $[W]_{ij}$ denotes the element at ij-th place in the matrix of W with respect to the standard bases of $\mathbb{R}^k$ and $\mathbb{R}^l$.

Let $C_b(\mathbb{R})$ be the space of continuous and bounded functions and $C^k_b(\mathbb{R})$ denotes the subset of at least k-times continuously differentiable functions. The norm on $C_b(\mathbb{R})$ is given by $\|\sigma\|_\infty := \sup_{z\in\mathbb{R}} |\sigma(z)|$. Moreover, for a function $\eta : \mathbb{R} \to \mathbb{R}$, let us define the Lipschitz constant of η as

$$\|\eta\|_{\mathrm{Lip}} = \inf\{L > 0 \mid \forall x, y \in \mathbb{R}\, |\eta(x) - \eta(y)| \leq L|x - y|\}.$$

The set of those $\mathbb{R} \to \mathbb{R}$ functions for which $\|.\|_{\text{Lip}}$ is finite is denoted by $\text{Lip}(\mathbb{R})$. In the sequel, we employ the convention that $\sum_k^l = 0$ and $\prod_k^l = 1$ whenever $k, l \in \mathbb{Z}$, $k > l$.

Let us fix a function $\sigma : \mathbb{R} \to \mathbb{R}$ to serve as the activation function of our neural network. We assume that $\sigma \in C_b^1(\mathbb{R})$ and $\sigma' \in C_b(\mathbb{R}) \cap \text{Lip}(\mathbb{R})$. Note that these assumptions imply the Lipschitz-continuity of σ, too. The Sobolev space $W^{1,\infty}(\mathbb{R})$ is just the space of Lipschitz functions moreover the norm on this space is $\|\cdot\|_{1,\infty} = \|\cdot\|_\infty + \|\cdot\|_{\text{Lip}}$, therefore $\sigma' \in W^{1,\infty}(\mathbb{R})$ and it is natural to regard σ as an element of $\sigma \in W^{2,\infty}(\mathbb{R})$. The norm which we use frequently in the sequel is the $W^{2,\infty}(\mathbb{R})$-norm of σ that is

$$|||\sigma||| := \|\sigma\|_{2,\infty} = \|\sigma\|_\infty + \left\|\sigma'\right\|_\infty + \left\|\sigma'\right\|_{\text{Lip}}.$$

Next, we consider networks consisting of $n \in \mathbb{N}_+$ hidden layers, where the number of nodes in each layer is given by $(d_1, \ldots, d_n) \in \mathbb{N}_+^n$. The space of the learning parameters is

$$\mathbb{R}^d \cong \Theta := (\mathbb{R}^{d_n})^* \oplus \bigoplus_{i=1}^n \text{Lin}\left(\mathbb{R}^{d_{i-1}}, \mathbb{R}^{d_i}\right),$$

where $d := \dim(\Theta) = d_n + \sum_{i=1}^n d_i d_{i-1}$ and $d_0 = m - 1$ for some $m > 1$ which corresponds to the dimension of the training data sequence. For the diameter of the network, we introduce the notation

$$D := \max_{0 \le i \le n} d_j.$$

A general element of Θ is of the form $\theta = (\phi, \mathbf{w})$, where $\phi \in (\mathbb{R}^{d_n})^*$ is a linear functional aggregating the node's output and $\mathbf{w} := (W_1, W_2, \ldots, W_n)$ is the sequence of weight matrices, where $W_i \in \text{Lin}\left(\mathbb{R}^{d_{i-1}}, \mathbb{R}^{d_i}\right)$, $i = 1, \ldots, n$. The Euclidean norm on Θ is

$$|(\phi, \mathbf{w})| = \left(|\phi|^2 + \sum_{i=1}^n |W_i|^2\right)^{1/2}.$$

Let us further introduce the notations

$$\sigma(\mathbf{w}_i^j, \cdot) = \begin{cases} \sigma_{W_j} \circ \sigma_{W_{j-1}} \circ \ldots \circ \sigma_{W_i}(\cdot) & \text{if } 1 \le i \le j \le n \\ \text{id}_{\mathbb{R}^{d_j}} & \text{otherwise,} \end{cases}$$

where $\sigma_{W_i} : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ is a nonlinear map given by $[\sigma_{W_i}(z)]_l = \sigma([W_i z]_l)$, $z \in \mathbb{R}^{d_{i-1}}$, $l = 1, \ldots, d_i$, $i = 1, \ldots, n$.

Let $\mathbf{z} := (z_1, \ldots, z_{d_0}) \in \mathbb{R}^{m-1}$ represent an input vector. With this, the function computed by a neural network with the above characteristics is given by $f : \Theta \times \mathbb{R}^{m-1} \to \mathbb{R}$

$$f((\phi, \mathbf{w}), \mathbf{z}) := \phi(\sigma(\mathbf{w}_1^n, \mathbf{z})) \tag{16}$$

For all $r > 0$ and $\eta > 0$, we define the regularized empirical risk function $U : \Theta \times \mathbb{R}^m \to [0, \infty)$ such that

$$U(\theta, x) := (y - f(\theta, \mathbf{z}))^2 + \frac{\eta}{2(r+1)}|\theta|^{2(r+1)}, \tag{17}$$

where we used the simpler notation for the input $x := (\mathbf{z}, y)$. The second term in (17) serves to regularize the optimization problem. We seek to optimize the parameter θ in such a way that, for some $r > 0$ and $\eta > 0$, $\theta \mapsto u(\theta) := E[U(\theta, X)]$ is minimized where $X = (\mathbf{Z}, Y) \in \mathbb{R}^m$ is a pair of random variables, $\mathbf{Z}$ representing the input and Y the target. The target variable Y is assumed one-dimensional for simplicity. For the derivative of U with respect to the learning parameter, the following notation is used

$$H(\theta, x) := \partial_\theta U(\theta, x) = -2(y - f(\theta, \mathbf{z}))\partial_\theta f(\theta, \mathbf{z}) + \eta|\theta|^{2r}\theta, \tag{18}$$

Mode	Regularization	Taming
1.	insufficient	off
2.	sufficient	off
3.	insufficient	on
4.	sufficient	on

Table 1: Simulation modes.

Mode	$\overline{\log(U)}$	$\overline{\log(\Delta\theta)}$	$\overline{A}$
1.	∞	∞	0.0572
2.	∞	∞	0.0010
3.	∞	∞	0.3308
4.	1.0543	-1.9518	0.1907

Table 2: Simulation results.

where we refer to the first term in the sequel as $G : \Theta \times \mathbb{R}^m \to \Theta^* \cong \mathbb{R}^d$. Thus,

$$G(\theta, x) := -2(y - f(\theta, \mathbf{z}))\partial_\theta f(\theta, \mathbf{z}). \tag{19}$$

Further, it is shown that within the framework of (17) and (18), Assumptions 1 and 2 hold.

Proposition 4.1. *Assumption 1 is satisfied by G, which is given in (19). In particular,*

$|G(\theta, x) - G(\theta', x)| \leq L_1(1 + |x|)^\rho(1 + |\theta| + |\theta'|)^{q-1}|\theta - \theta'|$, *for all* $x \in \mathbb{R}^m$ *and* $\theta, \theta' \in \mathbb{R}^d$,

where $L_1 = 16(n + 1)D^{3/2}(1 + |||\sigma|||)^{2n+4}$, $\rho = 3$ *and* $q - 1 = 2n + 1$.

Remark 4.2. *Assumption 2 is trivially satisfied in the context of neural networks when X_0 has either bounded support or a distribution with enough bounded moments. Similarly, the initialization of the algorithm is chosen appropriately either by using deterministic values or samples from distributions with enough bounded moments.*

Thus, the main results of this paper, namely Theorem 2.9, Corollary 2.10 and, most importantly, Theorem 2.11 hold true in this setting.

5 Numerical simulations

We created and trained a single-layer neural network ($n = 1$) with $d_1 = 16$ nodes to approximate the identity function $x \mapsto x$ on $[-10, 10] \cap \mathbb{Z}$. For the activation function, we chose the following smooth approximation of the popular ReLU function.

$$\sigma(z) = \mathbf{1}_{z\leq 0} 0.01 \log(1 + e^{100z}) + \mathbf{1}_{z>0}\left[z + 0.01\log(1 + e^{-100z})\right] \text{ and } \sigma'(z) = \frac{1}{1 + e^{-100z}}$$

For training set $(x_n) = (z_n, y_n)$, $n = 1, \dots, N$, we generated $N = 3200$ random numbers drawn from the uniform distribution on $[-10, 10] \cap \mathbb{Z}$ such that $y_n = z_n$, $n = 1, \dots, N$. After trying manually some (10-20) hyperparameter configurations, we set $\eta = 10^{-6}$, $\beta = 10$, $\lambda = 3.25 \times 10^{-3}$ and performed simulations in four different modes (See Table 1). The average run-time on a Dell-Latitude 7490 Intel Core i7-8650U CPU @ 1.90GHz $\times$ 8 laptop is about 60-80 sec, where one simulation run could only make use of one CPU core. In non-tamed cases we iterated using the usual SGLD scheme $\theta^\lambda_{n+1} := \theta^\lambda_n - \lambda H(\theta^\lambda_n, x_{n+1}) + (2\lambda\beta^{-1})^{1/2}\xi_{n+1}$, for $n = 1, \dots, N$, while in the tamed mode, we used the newly introduced TUSLA (4), where ξ_n's and $\theta^\lambda_0 := \theta_0$ are i.i.d. standard $2d_1$-dimensional Gaussian random vectors. Note that the growth estimate (2) is satisfied whenever $r \geq 2$. Therefore, we say that the regularization is insufficient if $0 \leq r < 2$ and sufficient if $r \geq 2$ c.f. Table 1. After each step, the log-utility $\log(U(\theta_n, x_n))$ is calculated along with $\log((\Delta\theta)_n) = \log(|\theta_n - \theta_{n-1}|)$, $n = 1, \dots, N$. We evaluated the accuracy of our network after each training step on test data consisting of $M = 800$ random input number and the corresponding expected output $(x'_n) = (z'_n, y'_n)$, $n = 1, \dots, M$. The training accuracy after k iterations is $A_k = \frac{1}{M}\sum_{n=1}^M \mathbf{1}_{\{|f(\theta^\lambda_k, z'_n) - y'_n| < 1/2\}}$. We present the time average of log-utility, $\log(\Delta\theta)$ and A in Table 2. Log-utility and $\log(\Delta\theta)$ values (Figure 1 and 2) show that the iteration is stable only when the regularization is sufficient and the taming is on. In case of non-tamed schemes, utility function values rapidly diverge after small number of steps. Mode 3 scheme with insufficient regularization

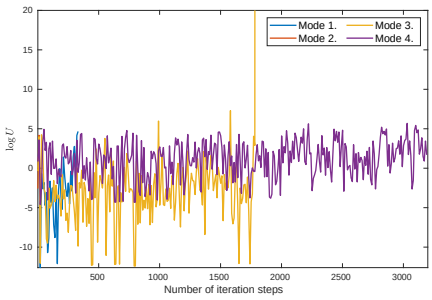

Figure 1: Log-utility values.

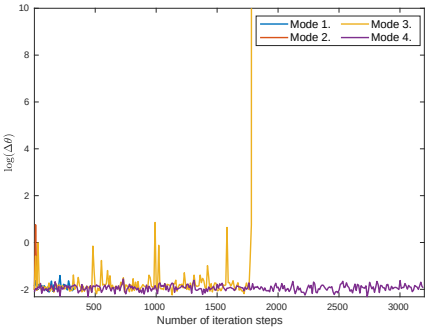

Figure 2: Log-parameter change values.

seems to converge, however as the utility values demonstrate, this scheme exhibits unstable behavior after large enough number of steps.

Figure 3 shows that the accuracy of non-tamed schemes reduced after small number of epochs.

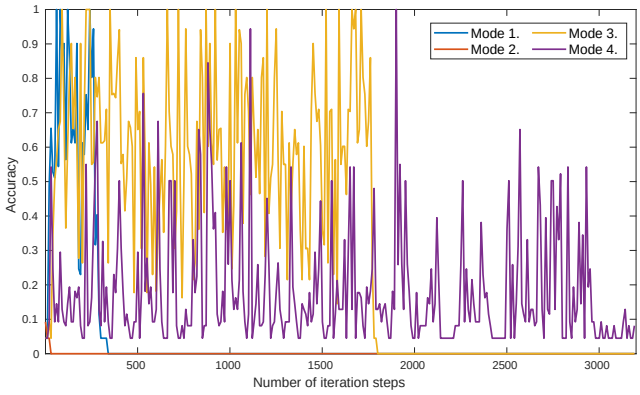

Figure 3: Training accuracy.

The Mode 3 scheme with insufficient regularization performs out any other variant up to the point when it becomes divergent. Regularization and taming together keep the learning parameter in a compact set. However, as we can see in 3, this does not guarantee that TUSLA finds and then remains at the global optimum. It visits rather local optima within this compact set and because of this, the training accuracy in Mode 4 fluctuates between 0 and 1. To overcome this issue, we can add memory to our network and store θ-values belonging to the maximal accuracy achieved during the learning process.

6 Conclusions

We introduce a new sampling algorithm, namely TUSLA (4), which can be used within the context of empirical risk minimization for neural networks. It does not have the stability shortcomings of other SGLD algorithms and our experiments demonstrate this important discovery. We also provide nonasymptotic estimates for TUSLA which explicitly bound the error between the target measure

and its law in Wasserstein-1 and 2 distances. Convergence rates and explicit constants are provided too.

References

M. Barkhagen, N. Chau, É. Moulines, M. Rásonyi, S. Sabanis, and Y. Zhang. On stochastic gradient Langevin dynamics with dependent data streams in the logconcave case. *To appear, Bernoulli, arXiv:1812.02709*, 2018.

N. Brosse, A. Durmus, and E. Moulines. The promises and pitfalls of stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, pages 8268–8278, 2018.

N. Brosse, A. Durmus, É. Moulines, and S. Sabanis. The tamed unadjusted Langevin algorithm. *Stochastic Processes and their Applications*, 129(10):3638–3663, 2019.

N. H. Chau, É. Moulines, M. Rásonyi, S. Sabanis, and Y. Zhang. On stochastic gradient Langevin dynamics with dependent data streams: the fully non-convex case. *arXiv preprint arXiv:1905.13142*, 2019.

X. Cheng, N. S. Chatterji, Y. Abbasi-Yadkori, P. L. Bartlett, and M. I. Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018.

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

A. S. Dalalyan and A. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

A. Durmus and E. Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

A. Eberle, A. Guillin, and R. Zimmer. Quantitative Harris-type theorems for diffusions and McKean–Vlasov processes. *Transactions of the American Mathematical Society*, 371(10):7135–7173, 2019a.

A. Eberle, A. Guillin, and R. Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *The Annals of Probability*, 47(4):1982–2010, 2019b.

M. Hutzenthaler, A. Jentzen, and P. E. Kloeden. Strong and weak divergence in finite time of euler's method for stochastic differential equations with non-globally lipschitz continuous coefficients. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 467(2130):1563–1576, 2011. ISSN 1364-5021.

M. Hutzenthaler, A. Jentzen, and P. E. Kloeden. Strong convergence of an explicit numerical method for sdes with nonglobally lipschitz continuous coefficients. *Ann. Appl. Probab.*, 22(4):1611–1641, 08 2012.

C.-R. Hwang. Laplace's method revisited: weak convergence of probability measures. *The Annals of Probability*, 8(6):1177–1182, 1980.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703, 2017.

S. Sabanis. A note on tamed euler approximations. *Electron. Commun. Probab.*, 18(47):1–10, 2013.

S. Sabanis. Euler approximations with varying coefficients: the case of superlinearly growing diffusion coefficients. *Ann. Appl. Probab.*, 26(4):2083–2105, 2016.

S. Sabanis and Y. Zhang. Higher order Langevin Monte Carlo algorithm. *Electronic Journal of Statistics*, 13(2):3805–3850, 2019.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, 2011.

P. Xu, J. Chen, D. Zou, and Q. Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3122–3133, 2018.

Y. Zhang, Ö. D. Akyildiz, T. Damoulas, and S. Sabanis. Nonasymptotic estimates for Stochastic Gradient Langevin Dynamics under local conditions in nonconvex optimization. *arXiv preprint arXiv:1910.02008*, 2019.

A Proofs

A.1 Complementary details to Section 2.1

Remark A.1. *By Assumption 1, since the function*

$$\phi_{i,h} = \frac{|G(\theta, X_0) - G(\theta + he_i, X_0)|}{h}$$

can be dominated for all $i = 1, \dots d$,$h < 1$ by the random variable $Z = L_1(1 + |X_0|)^{\rho}(2 + 2|\theta|)^{q-1}$ and $\mathbb{E}(Z) < \infty$, using a dominated convergence argument it can be concluded that partial derivation and expectation can be interchanged. As a result, $g \in C^1$ and consequently $h \in C^1$.

Constants in Remark 2.4. *One observes that it suffices to show*

$$\eta|\theta|^{2r+2} - \mathbb{E}[K(X_0)]|\theta|(1 + |\theta|^q) \geq A|\theta|^2 - B \tag{20}$$

for some suitable A and B or, equivalently,

$$\eta|\theta|^{2r+2} + B \geq A|\theta|^2 + \mathbb{E}[K(X_0)]|\theta|(1 + |\theta|^q).$$

Thus, setting $A = \mathbb{E}[K(X_0)]$ yields that (20) is satisfied with $B = (3\,\mathbb{E}[K(X_0)])^{q+2}\,\eta^{-q-1}$.

Proof of Proposition 2.6. Denote H_g the Hessian with respect to the antiderivative of g and H_{reg} the Hessian of the antiderivative of the regularization part. Then, the Hessian with respect to the antiderivative of h is

$$H_h = H_{reg} + H_g.$$

First of all, from the polynomial Lipchitzness of g for all x

$$\frac{|\nabla g(x + he_i) - \nabla g(x)|}{h} \leq L_2(1 + |x| + |x + he_i|)^{(q-1)}$$

where $L_2 = L_1\,\mathbb{E}(1 + |X_0|)^{\rho}$.
This implies that

$$\sqrt{\sum_{j=1}^{d} \frac{\left(\frac{\partial g}{\partial x_j}(x + he_i) - \frac{\partial g}{\partial x_j}(x)\right)^2}{h^2}} \leq L_2(1 + |x| + |x + he_i|)^{(q-1)}.$$

As $h \to 0$, there follows

$$|H_g(x)e_i| \leq L_2(1 + 2|x|)^{(q-1)} \quad \forall\, i = 1, \dots d.$$

Let $u \in \mathbb{R}^d$ .Then, $u = \sum_{i=1}^{d} a_i e_i$. As a result,

$$|H_g(x)u| \leq \sum |a_i||H_g(x)e_i| \leq L_2(1 + 2|x|)^{(q-1)} \sum |a_i| \leq \sqrt{d}L_2(1 + 2|x|)^{(q-1)}|u|.$$

Since u was arbitrary,

$$||H_g(x)||_2 \leq \sqrt{d}L_2(1 + 2|x|)^{(q-1)}.$$

The Hessian is symmetric which means that for all eigenvalues we have

$$\lambda + \sqrt{d}L_2(1 + 2|x|)^{(q-1)} \geq 0$$

so the matrix $A(x) = H_g(x) + \sqrt{d}L_2(1 + 2|x|)^{(q-1)} I_d$ is semi-positive definite. After some simple calculations one deduces that

$$H_{reg}(x) = \eta|x|^{2r} I_d + \eta 4r|x|^{2r-1} xx^T. \tag{21}$$

where it is observed that the second term is semi-positive definite. Let

$$R = \max\{(2^{3(q-1)+1}\sqrt{d}\frac{L_2}{\eta})^{\frac{1}{2r-q}}, (2^q\sqrt{d}\frac{L_2}{\eta})^{\frac{1}{2r}}.\} \tag{22}$$

There exists a constant a such that for all $|x| > R$ the quantity

$$\eta|x|^{2r} - \sqrt{d}L_2(1 + 2|x|)^{(q-1)} > 0$$

which yields that

$$\eta|x|^{2r} - \sqrt{d}L_2(1 + 2|x|)^{(q-1)} + \sqrt{d}L_2(1 + 2|R|)^{(q-1)} > 0, \quad \forall\, x : \quad |x| > R.$$

On the other hand, if $|x| \leq R$ one obtains

$$\eta|x|^{2r} - \sqrt{d}L_2(1 + 2|x|)^{(q-1)} + \sqrt{d}L_2(1 + 2|R|)^{(q-1)} \geq 0.$$

Thus, one concludes that for all $x \in \mathbb{R}^d$, the matrix

$$B(x) = \eta|x|^{2r} I_d - \sqrt{d}L_2(1 + 2|x|)^{(q-1)} I_d + \sqrt{d}L_2(1 + 2|R|)^{(q-1)} I_d$$

is positive definite. As a result, the matrix $A(x) + B(x) + \eta 4r|x|^{2r-1}xx^T = H_{reg} + H_g + \sqrt{d}L_2(1 + 2|R|)^{(q-1)} I_d$ is positive definite, which yields

$$\langle \theta - \theta', h(\theta) - h(\theta') \rangle \geq -a|\theta - \theta'|^2,$$

where $a = \sqrt{d}L_2(1 + 2|R|)^{q-1}$. □

Proof of Proposition 2.7. Let the reguralisation part $\Theta(\theta) := \eta\theta|\theta|^{2r}$ for any $\theta \in \mathbb{R}^d$. By using the mean value theorem, one deduces

$$|\Theta(\theta) - \Theta(\theta')| \leq ||H_{reg}(t\theta + (1-t)\theta')||_2 |\theta - \theta'|, \quad \text{for some } t \in [0,1],$$

where $||\cdot||_2$ denotes the spectral norm of a matrix. Due to (21), one observes that

$$||H_{reg}(x)||_2 \leq \eta|x|^{2r} + \eta 4r|x|^{2r+1} \leq 4r\eta(1+2|x|^{2r+1}) \leq 8r\eta(1+|x|)^{2r+1}.$$

Thus,

$$|\Theta(\theta) - \Theta(\theta')| \leq 8r\eta(1+|t\theta + (1-t)\theta'|)^{2r+1}|\theta - \theta'| \leq 8r\eta(1+|\theta| + |\theta'|)^{2r+1}|\theta - \theta'|.$$

In view of Assumption 1, the desired result follows. □

A.2 Complementary details to Section 3

Lemma A.2. *Let Assumptions 1 and 2 hold. Then, for any λ such that $0 < \lambda \leq \lambda_{max}$, one obtains for every $n \in \mathbb{N}$,*

$$\mathbb{E}\,|\theta_{n+1}|^2 \leq (1 - \frac{\eta}{2}\sqrt{\lambda})^n\, \mathbb{E}\,|\theta_0|^2 + 2\left(C_X \eta^{-1} + 2M_0^2(2+\eta) + 2d(\eta\beta)^{-1}\sqrt{\lambda_{max}}\right)$$

and, moreover,

$$\sup_n E|\theta_n^\lambda|^2 < \mathbb{E}\,|\theta_0|^2 + 2\left(C_X \eta^{-1} + 2M_0^2(2+\eta) + 2d(\eta\beta)^{-1}\sqrt{\lambda_{max}}\right),$$

where C_X is given in (27) and M_0 in the proof.

Proof. One first observes that, due to (7) and (8),

$$\begin{aligned}
&2\lambda\,\mathbb{E}\left[\langle \frac{\theta_n^\lambda}{|\theta_n^\lambda|^2}, H_\lambda(\theta_n^\lambda, X_{n+1})\rangle - \frac{\lambda}{2|\theta_n^\lambda|^2}|H_\lambda(\theta_n^\lambda, X_{n+1})|^2 |\theta_n^\lambda\right] \\
&\geq 2\lambda\,\mathbb{E}\left[\langle \frac{\theta_n^\lambda}{|\theta_n^\lambda|^2}, \frac{G(\theta_n^\lambda, X_{n+1}) + \eta\theta_n^\lambda|\theta_n^\lambda|^{2r}}{1+\sqrt{\lambda}|\theta_n^\lambda|^{2r}}\rangle - \frac{\lambda}{2|\theta_n^\lambda|^2}|H_\lambda(\theta_n^\lambda, X_{n+1})|^2 |\theta_n^\lambda\right] \\
&= 2\lambda \frac{1}{|\theta_n^\lambda|^2(1+\sqrt{\lambda}|\theta_n^\lambda|^{2r})}\left(\langle \theta_n^\lambda, \mathbb{E}\,G(\theta_n^\lambda, X_0)\rangle + \eta|\theta_n^\lambda|^{2r+2}\right) - 2\lambda\frac{4\,\mathbb{E}[K^2(X_0)]}{|\theta_n^\lambda|^2} - 2\lambda\eta^2 \\
&\geq 2\lambda\left(\frac{-\mathbb{E}(K(X_0)(|\theta_n^\lambda| + |\theta_n^\lambda|^{q+1})}{|\theta_n^\lambda|^2(1+\sqrt{\lambda}|\theta_n^\lambda|^{2r})} + \frac{\eta|\theta_n^\lambda|^{2r}}{1+\sqrt{\lambda}|\theta_n^\lambda|^{2r}} - \frac{4\,\mathbb{E}[K^2(X_0)]}{|\theta_n^\lambda|^2} - \eta^2\right). \qquad (23)
\end{aligned}$$

Since the function

$$f(\theta) := \left(\frac{-\mathbb{E}(K(X_0)(|\theta| + |\theta|^{q+1})}{|\theta|^2(1+\sqrt{\lambda}|\theta|^{2r})} + \frac{\eta|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}} - \frac{4\,\mathbb{E}[K^2(X_0)]}{|\theta|^2} - \eta^2\right) \qquad (24)$$

tends to $\frac{\eta}{\sqrt{\lambda}} - \eta^2$ as $|\theta| \to \infty$, it follows that there exists $M_0 > 0$ such that

$$|\theta_n^\lambda| \geq M_0 \implies f(\theta_n^\lambda) \geq \frac{1}{2}(\frac{\eta}{\sqrt{\lambda}} - \eta^2) = \frac{\eta}{2\sqrt{\lambda}}(1 - \sqrt{\lambda}\eta),$$

Then, as $\lambda \leq \frac{1}{4\eta^2}$,

$$|\theta_n^\lambda| \geq M_0 \implies f(\theta_n^\lambda) \geq \frac{\eta}{4\sqrt{\lambda}} \qquad (25)$$

Furthermore, for every $M > 0$, let us define

$$A_{n,M} := \{\omega \in \Omega : |\theta_n^\lambda| \geq M\}.$$

Combining (23), (24) and (25), yields that

$$\mathbb{E}\left[\lambda\left(2\langle\theta_n^\lambda, H_\lambda(\theta_n, X_{n+1})\rangle - \lambda|H_\lambda(\theta_n, X_{n+1})|^2\right)\mathbf{1}_{\mathrm{A_{n,M_0}}}\,|\theta_n^\lambda\right] \geq \frac{\eta}{2}\sqrt{\lambda}|\theta_n^\lambda|^2\,\mathbf{1}_{\mathrm{A_{n,M_0}}}. \qquad (26)$$

Thus,

$$\begin{aligned}
\mathbb{E}[|\theta_{n+1}|^2\,\mathbf{1}_{\mathrm{A_{n,M_0}}}\,|\theta_n^\lambda] &= \mathbb{E}\left[\left(|\theta_n^\lambda|^2 - 2\lambda\langle\theta_n^\lambda, H_\lambda(\theta_n^\lambda, X_{n+1})\rangle + \lambda^2|H_\lambda(\theta_n^\lambda, X_{n+1})|^2 + \frac{2\lambda}{\beta}|\xi_{n+1}|^2\right)\mathbf{1}_{\mathrm{A_{n,M_0}}}\,|\theta_n^\lambda\right] \\
&\leq (1 - \frac{\eta}{2}\sqrt{\lambda})|\theta_n^\lambda|^2\,\mathbf{1}_{\mathrm{A_{n,M_0}}} + 2\frac{\lambda}{\beta}d\,\mathbf{1}_{\mathrm{A_{n,M_0}}} \\
&< (1 - \frac{\eta}{2}\sqrt{\lambda})|\theta_n^\lambda|^2\,\mathbf{1}_{\mathrm{A_{n,M_0}}} + \sqrt{\lambda}\left(C_X + 2M_0^2(2\eta + \eta^2) + 2d\beta^{-1}\sqrt{\lambda_{max}}\right)\mathbf{1}_{\mathrm{A_{n,M_0}}},
\end{aligned}$$

where

$$C_X := 2\,\mathbb{E}[K(X_0)]M_0 + 4\,\mathbb{E}[K^2(X_0)]. \qquad (27)$$

On the other hand, due to (7) and (8),

$$\mathbb{E}\left[|\theta_{n+1}|^2\,\mathbf{1}_{\mathrm{A^C_{n,M_0}}}\,|\theta_n^\lambda\right] = \left(|\theta_n^\lambda|^2 - \mathbb{E}\left[\lambda\left(2\langle\theta_n^\lambda, H_\lambda(\theta_n^\lambda, X_{n+1})\rangle - \lambda|H_\lambda(\theta_n^\lambda, X_{n+1})|^2\right)|\theta_n^\lambda\right] + 2\frac{\lambda}{\beta}d\right)\mathbf{1}_{\mathrm{A^C_{n,M_0}}}$$

$$
\begin{aligned}
&\leq \left(|\theta_n^\lambda|^2 + 2\lambda|\theta_n^\lambda|\,\mathbb{E}\left[|H_\lambda(\theta_n^\lambda, X_{n+1})|\theta_n^\lambda\right] + \lambda^2\,\mathbb{E}\left[|H_\lambda(\theta_n^\lambda, X_{n+1})|^2|\theta_n^\lambda\right] + 2\frac{\lambda}{\beta}d\right)\mathbf{1}_{\mathrm{A}^{\mathrm{C}}_{\mathrm{n,M_0}}}\\
&< \left(|\theta_n^\lambda|^2 + \sqrt{\lambda}\left(C_X + 2M_0^2(\eta+\eta^2) + 2d\beta^{-1}\sqrt{\lambda_{max}}\right)\right)\mathbf{1}_{\mathrm{A}^{\mathrm{C}}_{\mathrm{n,M_0}}}\\
&= (1-\frac{\eta}{2}\sqrt{\lambda})|\theta_n^\lambda|^2\,\mathbf{1}_{\mathrm{A}^{\mathrm{C}}_{\mathrm{n,M_0}}} + \sqrt{\lambda}\left(\frac{\eta}{2}|\theta_n^\lambda|^2 + C_X + 2M_0^2(\eta+\eta^2) + 2d\beta^{-1}\sqrt{\lambda_{max}}\right)\mathbf{1}_{\mathrm{A}^{\mathrm{C}}_{\mathrm{n,M_0}}}\\
&< (1-\frac{\eta}{2}\sqrt{\lambda})|\theta_n^\lambda|^2\,\mathbf{1}_{\mathrm{A}^{\mathrm{C}}_{\mathrm{n,M_0}}} + \sqrt{\lambda}\left(C_X + 2M_0^2(2\eta+\eta^2) + 2d\beta^{-1}\sqrt{\lambda_{max}}\right)\mathbf{1}_{\mathrm{A}^{\mathrm{C}}_{\mathrm{n,M_0}}}.
\end{aligned}
$$

Combining the two estimates above yields

$$
\mathbb{E}\left[|\theta_{n+1}|^2|\theta_n^\lambda\right] < (1-\frac{\eta}{2}\sqrt{\lambda})|\theta_n^\lambda|^2 + \sqrt{\lambda}\left(C_X + 2M_0^2(2\eta+\eta^2) + 2d\beta^{-1}\sqrt{\lambda_{max}}\right)
$$

which implies

$$
\begin{aligned}
\mathbb{E}\,|\theta_{n+1}|^2 <&(1-\frac{\eta}{2}\sqrt{\lambda})^n\,\mathbb{E}\,|\theta_0|^2 + \sqrt{\lambda}\left(C_X + 2M_0^2(2\eta+\eta^2) + 2d\beta^{-1}\sqrt{\lambda_{max}}\right)\sum_{j=0}^{\infty}(1-\frac{\eta}{2}\sqrt{\lambda})^j\\
\leq&(1-\frac{\eta}{2}\sqrt{\lambda})^n\,\mathbb{E}\,|\theta_0|^2 + 2\left(C_X\eta^{-1} + 2M_0^2(2+\eta) + 2d(\eta\beta)^{-1}\sqrt{\lambda_{max}}\right)
\end{aligned}
$$

that gives the desired result. □

Proof of Lemma 3.1. First one defines, for every $n \in \mathbb{N}$,

$$
\Delta_n := \theta_n^\lambda - \lambda H_\lambda(\theta_n^\lambda, X_{n+1}). \tag{28}
$$

Then, one calculates that, for any integer $p > 1$ (since the case $p = 1$ is covered by Lemma A.2),

$$
|\theta_{n+1}^\lambda|^{2p} = \left(|\Delta_n|^2 + \frac{2\lambda}{\beta}|\xi_{n+1}|^2 + 2\langle\Delta_n, \sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\rangle\right)^p.
$$

Hence,

$$
\begin{aligned}
\mathbb{E}\left[|\theta_{n+1}^\lambda|^{2p}|\theta_n^\lambda\right] &= \mathbb{E}\left[\left(|\Delta_n|^2 + \frac{2\lambda}{\beta}|\xi_{n+1}|^2 + 2\langle\Delta_n, \sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\rangle\right)^p|\theta_n^\lambda\right]\\
&= \sum_{k_1+k_2+k_3=p}\frac{p!}{k_1!k_2!k_3!}\,\mathbb{E}\left[|\Delta_n|^{2k_1}\left|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\right|^{2k_2}\left(2\langle\Delta_n, \sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\rangle\right)^{k_3}|\theta_n^\lambda\right]\\
&\leq \mathbb{E}[|\Delta_n|^{2p}|\theta_n^\lambda] + 2p\,\mathbb{E}\left[|\Delta_n|^{2p-2}\langle\Delta_n, \sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\rangle|\theta_n^\lambda\right]\\
&\qquad + \sum_{k=2}^{2p}\binom{2p}{k}\,\mathbb{E}\left[|\Delta_n|^{2p-k}\left|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\right|^k|\theta_n^\lambda\right]\\
&\leq \mathbb{E}[|\Delta_n|^{2p}|\theta_n^\lambda] + \mathbb{E}\left[\sum_{l=0}^{2(p-1)}\binom{2p}{l+2}\left(|\Delta_n|^{2(p-1)-l}\left|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\right|^{(q-1)}\right)\frac{2\lambda}{\beta}|\xi_{n+1}|^2|\theta_n^\lambda\right]\\
&= \mathbb{E}[|\Delta_n|^{2p}|\theta_n^\lambda] + \mathbb{E}\left[\binom{2p}{2}\sum_{l=0}^{2(p-1)}\binom{2(p-1)}{l}\left(|\Delta_n|^{2(p-1)-l}\left|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\right|^l\right)\frac{2\lambda}{\beta}|\xi_{n+1}|^2|\theta_n^\lambda\right]\\
&\leq \mathbb{E}[|\Delta_n|^{2p}|\theta_n^\lambda] + 2^{2p-3}p(2p-1)\,\mathbb{E}[|\Delta_n|^{2p-2}|\theta_n^\lambda]\frac{2\lambda}{\beta}d + 2^{2p-3}p(2p-1)\left(\frac{2\lambda}{\beta}\right)^p\mathbb{E}\,|\xi_{n+1}|^{2p}.
\end{aligned} \tag{29}
$$

Let us also define, for every $n \in \mathbb{N}$,

$$
r_n := -2\lambda\langle\theta_n^\lambda, H_\lambda(\theta_n^\lambda, X_{n+1})\rangle + \lambda^2|H_\lambda(\theta_n^\lambda, X_{n+1})|^2 \tag{30}
$$

and observe that, due to (28),

$$
|\Delta_n|^2 = |\theta_n^\lambda|^2 + r_n.
$$

Consequently,

$$
\begin{aligned}
\mathbb{E}[|\Delta_n|^{2p}|\theta_n^\lambda] &= \sum_{k=0}^{p}\binom{p}{k}|\theta_n^\lambda|^{2(p-k)}\,\mathbb{E}\left[r_n^k|\theta_n^\lambda\right]\\
&= |\theta_n^\lambda|^{2p} + p|\theta_n^\lambda|^{2p-2}\,\mathbb{E}[r_n|\theta_n^\lambda] + \sum_{k=2}^{p}\binom{p}{k}|\theta_n^\lambda|^{2(p-k)}\,\mathbb{E}\left[r_n^k|\theta_n^\lambda\right]
\end{aligned} \tag{31}
$$

Let us also define the constant M by the following expression

$$M := \max\{M_0, 1, \max_{2\le k\le p}\left(\binom{p}{k}\binom{k}{\lceil \frac{k}{2}\rceil}2^{4k}(1+\mathbb{E}[K^{2k}(X_0)])\frac{4(p+1)}{\eta}\right)^{\frac{1}{k}}, \\ \max_{2\le k\le p-1}\left(\binom{p-1}{k}\binom{k}{\lceil \frac{k}{2}\rceil}2^{4k}(1+\mathbb{E}[K^{2k}(X_0)])\frac{4p}{\eta}\right)^{\frac{1}{k}}, \sqrt{2^{2p-3}(2p-1)p\frac{d}{\beta\eta}}\}. \tag{32}$$

When $|\theta_n^\lambda| > M$ and due to the fact that $\lambda \le 1$, see (6), one obtains

$$\begin{aligned}
|\lambda\langle\theta_n^\lambda, H_\lambda(\theta_n^\lambda, X_{n+1})\rangle| &\le \frac{\lambda K(X_{n+1})(1+|\theta_n^\lambda|^q)|\theta_n^\lambda| + \lambda\eta|\theta_n^\lambda|^{2r+2}}{1+\sqrt{\lambda}|\theta_n^\lambda|^{2r}} \\
&\le \frac{\lambda K(X_{n+1})(|\theta_n^\lambda| + |\theta_n|^{q+1})}{1+\sqrt{\lambda}|\theta_n^\lambda|^{2r}} + \sqrt{\lambda}\eta|\theta_n^\lambda|^2 \\
&\le \frac{\lambda K(X_{n+1})(2+2|\theta_n^\lambda|^{2r})}{1+\sqrt{\lambda}|\theta_n^\lambda|^{2r}} + \sqrt{\lambda}\eta|\theta_n^\lambda|^2 \\
&\le \frac{2\sqrt{\lambda}K(X_{n+1})(\sqrt{\lambda}+\sqrt{\lambda}|\theta_n^\lambda|^{2r})}{1+\sqrt{\lambda}|\theta_n^\lambda|^{2r}} + \sqrt{\lambda}\eta|\theta_n^\lambda|^2 \\
&\le 2\sqrt{\lambda}K(X_{n+1}) + \sqrt{\lambda}\eta|\theta_n^\lambda|^2 \\
&\le (\sqrt{a_n}+\sqrt{b_n})|\theta_n^\lambda|
\end{aligned} \tag{33}$$

where $a_n = 2\sqrt{\lambda}K(X_{n+1})$ and $b_n = \sqrt{\lambda}\eta|\theta_n^\lambda|$ since $|\theta_n^\lambda| > M \ge 1$. In addition,

$$\begin{aligned}
|\lambda^2 H_\lambda^2(\theta_n^\lambda, X_{n+1})| &\le \frac{2\lambda^2K^2(X_{n+1})(1+|\theta_n^\lambda|^q)^2 + 2\lambda^2\eta^2|\theta_n^\lambda|^{4r+2}}{1+\lambda|\theta_n^\lambda|^{4r}} \\
&\le \frac{4\lambda^2K^2(X_{n+1})(1+|\theta_n^\lambda|^{2q})}{1+\lambda|\theta_n^\lambda|^{4r}} + 2\lambda\eta^2|\theta_n^\lambda|^2 \\
&\le \frac{4\lambda^2K^2(X_{n+1})(2+|\theta_n^\lambda|^{4r})}{1+\lambda|\theta_n^\lambda|^{4r}} + 2\lambda\eta^2|\theta_n^\lambda|^2 \\
&\le \frac{4\lambda K^2(X_{n+1})(2\lambda+\lambda|\theta_n^\lambda|^{4r})}{1+\lambda|\theta_n^\lambda|^{4r}} + 2\lambda\eta^2|\theta_n^\lambda|^2 \\
&\le 8\lambda K^2(X_{n+1}) + 2\lambda\eta^2|\theta_n^\lambda|^2 \\
&= 2a_n + 2b_n.
\end{aligned} \tag{34}$$

Observing that, due to (30), (33) and (34),

$$\begin{aligned}
r_n^k &= \sum_{j=0}^k\binom{k}{j}2^{k-j}(\sqrt{a_n}+\sqrt{b_n})^{k-j}2^j(a_n+b_n)^j|\theta_n^\lambda|^{k-j} \\
&= \sum_{j=0}^k\binom{k}{j}2^k((\sqrt{a_n}+\sqrt{b_n})^2)^{\frac{k-j}{2}}(a_n+b_n)^j|\theta_n^\lambda|^{k-j} \\
&= \sum_{j=0}^k\binom{k}{j}2^k2^{\frac{k-j}{2}}(a_n+b_n)^{\frac{k-j}{2}}(a_n+b_n)^j|\theta_n^\lambda|^{k-j} \\
&= \sum_{j=0}^k\binom{k}{j}2^{\frac{k+j}{2}}(a_n+b_n)^{\frac{k+j}{2}}|\theta_n^\lambda|^{k-j} \\
&\le \sum_{j=0}^k\binom{k}{j}2^{k+j}(a_n^{\frac{k+j}{2}}+b_n^{\frac{k+j}{2}})|\theta_n^\lambda|^{k-j}
\end{aligned}$$

yields that

$$\mathbb{E}[r_n^k|\theta_n^\lambda] \le \sum_{j=0}^k\binom{k}{j}2^{2(k+j)}\lambda^{\frac{k+j}{2}}\mathbb{E}[K^{k+j}(X_0)]|\theta_n^\lambda|^{k-j} + \sum_{j=0}^k\binom{k}{j}2^{k+j}\lambda^{\frac{k+j}{2}}\eta^{k+j}|\theta_n^\lambda|^{k+j}|\theta_n^\lambda|^{k-j}.$$

Consequently, and in view of (31),

$$\begin{aligned}
\mathbb{E}[|\Delta_n|^{2p}|\theta_n^\lambda] \le& |\theta_n^\lambda|^{2p} + p|\theta_n^\lambda|^{2p-2}\mathbb{E}[r_n|\theta_n^\lambda] + \sum_{k=2}^p\binom{p}{k}|\theta_n^\lambda|^{2p-2k} \\
&\times\left[\sum_{j=0}^k\binom{k}{j}2^{2(k+j)}\lambda^{\frac{k+j}{2}}\mathbb{E}[K^{k+j}(X_0)]|\theta_n^\lambda|^{k-j} + |\theta_n^\lambda|^{2k}\sum_{j=0}^k\binom{k}{j}2^{k+j}\lambda^{\frac{k+j}{2}}\eta^{k+j}.\right]
\end{aligned}$$

Moreover, due to (26),

$$p|\theta_n^\lambda|^{2p-2}\,\mathbb{E}[r_n|\theta_n^\lambda] \leq -p\frac{1}{2}\sqrt{\lambda}\eta|\theta_n^\lambda|^{2p},$$

and thus one obtains

$$\begin{aligned}
\mathbb{E}[|\Delta_n|^{2p}|\theta_n^\lambda] &\leq |\theta_n^\lambda|^{2p} - p\frac{1}{2}\sqrt{\lambda}\eta|\theta_n^\lambda|^{2p} + \sum_{k=2}^{p}\binom{p}{k}|\theta_n^\lambda|^{2p-2k}\left[\sum_{j=0}^{k}\binom{k}{j}2^{2(k+j)}\lambda^{\frac{k+j}{2}}\,\mathbb{E}[K^{k+j}(X_0)]\frac{|\theta_n^\lambda|^{2k}}{|\theta_n^\lambda|^{k+j}}\right] \\
&\qquad + \sum_{k=2}^{p}\binom{p}{k}|\theta_n^\lambda|^{2p}\left[\sum_{j=0}^{k}\binom{k}{j}2^{k+j}\lambda^{\frac{k+j}{2}}\eta^{k+j}\right] \\
&\leq |\theta_n^\lambda|^{2p} - \frac{1}{2}p\sqrt{\lambda}\eta|\theta_n^\lambda|^{2p} + \sum_{k=2}^{p}\binom{p}{k}|\theta_n^\lambda|^{2p}\left[\sum_{j=0}^{k}\binom{k}{j}2^{2(k+j)}\lambda^{\frac{k+j}{2}}\,\mathbb{E}[K^{k+j}(X_0)](\frac{1}{M})^{k+j}\right] \\
&\qquad + \sum_{k=2}^{p}\binom{p}{k}|\theta_n^\lambda|^{2p}\left[\sum_{j=0}^{k}\binom{k}{j}2^{k+j}\lambda^{\frac{k+j}{2}}\eta^{k+j}\right] \\
&\leq |\theta_n^\lambda|^{2p} - \frac{1}{2}p\sqrt{\lambda}\eta|\theta_n^\lambda|^{2p} \\
&+ |\theta_n^\lambda|^{2p}\sum_{k=2}^{p}\binom{p}{k}\left[\sum_{j=0}^{k}\binom{k}{j}2^{2(k+j)}\lambda^{\frac{k+j}{2}}\,\mathbb{E}\,K^{k+j}(X_0)\left(\frac{1}{M}\right)^k + \sum_{j=0}^{k}\binom{k}{j}2^{k+j}\lambda^{\frac{k+j}{2}}\eta^{k+j}\right]
\end{aligned}$$

Applying the previous relation for $p-1$ and bringing it all together using (29),

$$\begin{aligned}
\mathbb{E}[|\theta_{n+1}^\lambda|^{2p}|\theta_n^\lambda] &\leq |\theta_n^\lambda|^{2p} - \frac{1}{2}p\sqrt{\lambda}\eta|\theta_n^\lambda|^{2p} \\
&+ |\theta_n^\lambda|^{2p}\sum_{k=2}^{p}\binom{p}{k}\left[\sum_{j=0}^{k}\binom{k}{j}2^{2(k+j)}\lambda^{\frac{k+j}{2}}\,\mathbb{E}[K^{k+j}(X_0)]\left(\frac{1}{M}\right)^k + \sum_{j=0}^{k}\binom{k}{j}2^{k+j}\lambda^{\frac{k+j}{2}}\eta^{k+j}\right] \\
&+ 2^{2p-3}(2p-1)2\frac{\lambda}{\beta}d(|\theta_n^\lambda|^{2p-2} - \frac{1}{2}(p-2)\sqrt{\lambda}\eta|\theta_n^\lambda|^{2p-2} \\
&+ |\theta_n^\lambda|^{2p-2}\sum_{k=2}^{p-1}\binom{p-1}{k}\left(\sum_{j=0}^{k}\binom{k}{j}2^{2(k+j)}\lambda^{\frac{k+j}{2}}\,\mathbb{E}[K^{k+j}(X_0)]\left(\frac{1}{M}\right)^k + \sum_{j=0}^{k}\binom{k}{j}2^{k+j}\lambda^{\frac{k+j}{2}}\eta^{k+j}\right) \\
&+ 2^{2p-3}(2p-1)\left(\frac{2\lambda}{\beta}\right)^p\mathbb{E}\,|\xi_{n+1}|^{2p}. \qquad (35)
\end{aligned}$$

We now show that the restriction $\lambda \leq \min\{1, \frac{1}{4\eta^2\left(8(p+1)\left(\lceil\frac{p}{2}\rceil\right)^2\right)^2}\}$ yields the desired result. We start by showing that

$$\frac{(p-2)}{4}\sqrt{\lambda\eta} > \sum_{k=2}^{p}\binom{p}{k}[\sum_{j=0}^{k}\binom{k}{j}2^{2(k+j)}\lambda^{\frac{k+j}{2}}\,\mathbb{E}[K^{k+j}(X_0)]\left(\frac{1}{M}\right)^k.$$

Since for all $0 \leq j \leq k, 2 \leq k \leq p$

$$\begin{aligned}
\lambda^{\frac{k+j-1}{2}} &\leq 1 \\
&\leq \frac{\eta M^k}{4(p+1)\binom{p}{k}\binom{k}{\lceil\frac{k}{2}\rceil}2^{4k}(1+\mathbb{E}[K^{2k}(X_0)])} \\
&\leq \frac{\eta M^k}{4(p+1)\binom{p}{k}\binom{k}{j}2^{2(k+j)}\,\mathbb{E}[K^{k+j}(X_0)]},
\end{aligned}$$

one deduces that, for $0 \leq j \leq k, 2 \leq k \leq p$

$$\frac{\sqrt{\lambda}\eta}{4(p+1)} \geq \binom{p}{k}\binom{k}{j}2^{2(k+j)}\lambda^{\frac{k+j}{2}}\,\mathbb{E}[K^{k+j}(X_0)]\left(\frac{1}{M}\right)^k$$

which yields that

$$(p+1)\frac{\sqrt{\lambda}\eta}{4(p+1)} \geq (k+1)\frac{\sqrt{\lambda}\eta}{4(p+1)} \geq \binom{p}{k}\sum_{j=0}^{k}\binom{k}{j}2^{2(k+j)}\lambda^{\frac{k+j}{2}}\,\mathbb{E}[K^{k+j}(X_0)]\left(\frac{1}{M}\right)^k$$

Consequently,

$$\frac{1}{4}p\sqrt{\lambda}\eta \geq \frac{1}{2}\sqrt{\lambda}\eta + \frac{(p-2)}{4}\sqrt{\lambda}\eta \geq \frac{1}{2}\sqrt{\lambda}\eta + \sum_{k=2}^{p}\binom{p}{k}[\sum_{j=0}^{k}\binom{k}{j}2^{2(k+j)}\lambda^{\frac{k+j}{2}}\,\mathbb{E}\,K^{k+j}(X_0)\left(\frac{1}{M}\right)^k. \qquad (36)$$

Moreover, for $0 \leq j \leq k$ and $2 \leq k \leq p$,

$$\lambda \leq \frac{1}{4\eta^2 \left(8(p+1)\left(\lceil \frac{p}{2} \rceil\right)^2\right)^2} \leq \frac{1}{4\eta^2 \left(8(p+1)\left(\lceil \frac{p}{2} \rceil\right)^2\right)^{\frac{2}{k+j-1}}},$$

and thus,

$$\lambda^{\frac{k+j-1}{2}} \leq \frac{1}{4^{\frac{k+j-1}{2}} 8(p+1)\left(\lceil \frac{p}{2} \rceil\right)^2 \eta^{k+j-1}} = \frac{1}{2^{k+j} 4(p+1)\left(\lceil \frac{p}{2} \rceil\right)^2 \eta^{k+j-1}} \leq \frac{\eta}{4(p+1)\binom{k}{j}\binom{p}{k} 2^{k+j} \eta^{k+j}}$$

which leads to

$$\frac{p-2}{4}\sqrt{\lambda}\eta > \sum_{k=2}^{p} \binom{p}{k} \sum_{j=0}^{k} \binom{k}{j} 2^{k+j} \lambda^{\frac{k+j}{2}} \eta^{k+j}. \tag{37}$$

The combination of the inequalities (36), (37) yields

$$\begin{aligned}
\mathbb{E}[|\Delta_n|^{2p}|\theta_n^\lambda] \leq & |\theta_n^\lambda|^{2p} - \frac{1}{2}p\sqrt{\lambda}\eta|\theta_n^\lambda|^{2p} \\
& + |\theta_n^\lambda|^{2p} \sum_{k=2}^{p} \binom{p}{k} [\sum_{j=0}^{k} \binom{k}{j} 2^{2(k+j)} \lambda^{\frac{k+j}{2}} \mathbb{E}[K^{k+j}(X_0)] \left(\frac{1}{M}\right)^k + \sum_{j=0}^{k} \binom{k}{j} 2^{k+j} \lambda^{\frac{k+j}{2}} \eta^{k+j}] \\
\leq & (1 - \sqrt{\lambda}\eta)|\theta_n^\lambda|^{2p}.
\end{aligned} \tag{38}$$

Using similar arguments for $p-1$ leads to

$$\mathbb{E}[|\Delta_n|^{2p-2}|\theta_n^\lambda] \leq (1 - \sqrt{\lambda}\eta)|\theta_n^\lambda|^{2p-2} \leq \frac{1}{M^2}(1 - \sqrt{\lambda}\eta)|\theta_n^\lambda|^{2p} \tag{39}$$

Thus, when $|\theta_n^\lambda| \geq M$, and in view of (29), (38),(39) and (32), one obtains

$$\begin{aligned}
\mathbb{E}[|\theta_{n+1}^\lambda|^{2p} \mathbf{1}_{\mathrm{A_{n,M}}} |\theta_n^\lambda] \leq & (1 - \sqrt{\lambda}\eta)(1 + \frac{2^{2p-3}p(2p-1)\lambda d}{\beta M^2})|\theta_n^\lambda|^{2p} \mathbf{1}_{\mathrm{A_{n,M}}} \\
& + 2^{2p-3}(2p-1)p\left(\frac{2\lambda}{\beta}\right)^p \mathbb{E}\,|\xi_{n+1}|^{2p} \mathbf{1}_{\mathrm{A_{n,M}}} \\
\leq & (1 - \lambda\eta^2)|\theta_n^\lambda|^{2p} \mathbf{1}_{\mathrm{A_{n,M}}} + 2^{2p-3}(2p-1)p\left(\frac{2\lambda}{\beta}\right)^p \mathbb{E}\,|\xi_{n+1}|^{2p} \mathbf{1}_{\mathrm{A_{n,M}}}.
\end{aligned} \tag{40}$$

When $|\theta_n^\lambda| < M$, one observes that

$$\begin{aligned}
|\Delta_n|^{2p} &\leq |\theta_n^\lambda|^{2p} + \sum_{k=0}^{p-1} \binom{p}{k} |r_n|^{p-k} |\theta_n^\lambda|^{2k} \\
&\leq (1 - \lambda\eta^2)|\theta_n^\lambda|^{2p} + \lambda\eta^2 M^{2p} + \sum_{k=0}^{p-1} \binom{p}{k} 2^{p-k} M^{2k} (\lambda^{2(p-k)} |H_\lambda(\theta_n^\lambda, X_{n+1})|^{2(p-k)} \\
&\qquad + \lambda^{p-k} M^{2(p-k)} |H_\lambda(\theta_n^\lambda, X_{n+1})|^{p-k}).
\end{aligned}$$

Analysing the terms,

$$\begin{aligned}
\lambda^{p-k}|H_\lambda(\theta_n^\lambda, X_{n+1})|^{p-k} &\leq \lambda^{p-k}(K(X_{n+1})(1 + |\theta_n^\lambda|^q) + \eta|\theta_n^\lambda|^{2r+1})^{p-k} \\
&\leq 2^{p-k}\lambda^{p-k}(\lambda^{p-k}(K(X_{n+1})^{p-k}(1 + M^q)^{p-k} + \eta^{p-k} M^{(2r+1)(p-k)})
\end{aligned}$$

one obtains

$$\begin{aligned}
\mathbb{E}[|\Delta_n|^{2p} \mathbf{1}^{\mathrm{C}}_{\mathrm{A_{n,M}}} |\theta_n^\lambda] \leq & (1 - \lambda\eta^2)|\theta_n^\lambda|^{2p} \mathbf{1}^{\mathrm{C}}_{\mathrm{A_{n,M}}} + (\lambda\eta^2 M^{2p} \\
& + \lambda \sum_{k=0}^{p-1} \binom{p}{k} 2^{p-k} M^{2k} (R^2_{\lambda,M,p,\eta} + M^{2(p-k)} R_{M,p,\eta}) \mathbf{1}^{\mathrm{C}}_{\mathrm{A_{n,M}}} \\
= & (1 - \lambda\eta^2)|\theta_n^\lambda|^{2p} \mathbf{1}^{\mathrm{C}}_{\mathrm{A_{n,M}}} + \lambda\eta^2 M^{2p} + \lambda C(\eta, p, M) \mathbf{1}^{\mathrm{C}}_{\mathrm{A_{n,M}}}.
\end{aligned} \tag{41}$$

where

$$R_{M,p,\eta} = 2^{p-k}\left(\mathbb{E}[K(X_0)^{p-k}](1 + M^q)^{p-k} + \eta^{p-k} M^{(2r+1)(p-k)}\right)$$

and

$$C(\eta, p, M) = \sum_{k=0}^{p-1} \binom{p}{k} 2^{p-k} M^{2k} (R^2_{\lambda,M,p,\eta} + M^{2(p-k)} R_{M,p,\eta}).$$

Moreover, in a similar way to (41), one concludes that

$$\mathbb{E}[|\Delta_n|^{2p-2} \mathbf{1}^{\mathrm{C}}_{\mathrm{A_{n,M}}} |\theta_n^\lambda] \leq M^{2p-2} \mathbf{1}^{\mathrm{C}}_{\mathrm{A_{n,M}}} + \lambda C(\eta, p, M) \mathbf{1}^{\mathrm{C}}_{\mathrm{A_{n,M}}},$$

and hence

$$\begin{aligned}\mathbb{E}\left[|\theta_{n+1}^{\lambda}|^{2p}\mathbf{1}_{\mathrm{A_{n,M}^C}}|\theta_n^{\lambda}\right] \leq & (1-\lambda\eta^2)|\theta_n^{\lambda}|^{2p}\mathbf{1}_{\mathrm{A_{n,M}^C}} \\ & + \lambda\left(C(\eta,p,M)+\eta^2M^{2p}+2^{2p-3}p(2p-1)(C(\eta,p-1,M)+M^{2p-2})\frac{2}{\beta}d\right)\mathbf{1}_{\mathrm{A_{n,M}^C}} \\ & + \lambda\left(2^{2p-3}p(2p-1)\left(\frac{2}{\beta}\right)^p\mathbb{E}\left|\xi_{n+1}\right|^{2p}\right)\mathbf{1}_{\mathrm{A_{n,M}^C}}. \\ \leq & (1-\lambda\eta^2)|\theta_n^{\lambda}|^{2p}\mathbf{1}_{\mathrm{A_{n,M}^C}}+\lambda A_p\mathbf{1}_{\mathrm{A_{n,M}^C}} \end{aligned} \tag{42}$$

where

$$A_p = C(\eta,p,M)+\eta^2M^{2p}+2^{2p-3}p(2p-1)(C(\eta,p-1,M)+M^{2p-2})\frac{2}{\beta}d+2^{2p-3}p(2p-1)\left(\frac{2}{\beta}\right)^p\mathbb{E}\left|\xi_{n+1}\right|^{2p}. \tag{43}$$

Adding (40) and (42), one obtains

$$\begin{aligned}\mathbb{E}\left|\theta_{n+1}^{\lambda}\right|^{2p} \leq & (1-\lambda\eta^2)\,\mathbb{E}\left|\theta_n^{\lambda}\right|^{2p}+\lambda A_p \leq (1-\lambda\eta^2)^n\,\mathbb{E}\left|\theta_0\right|^{2p}+\frac{1}{\eta^2}A_p \\ \leq & (1-\lambda\eta^2)^n\,\mathbb{E}\left|\theta_0\right|^{2p}+C`_p\end{aligned}$$

where, in view of (43),

$$C`_p = \frac{1}{\eta^2}A_p \tag{44}$$

which yields the desired result. □

Proof of Lemma 3.5. This is an immediate consequence of Remark 3.3 and the definition of the Lyapunov function as given in (11) with $m=4$. □

Proof of Lemma 3.6. See (Chau et al., 2019, Lemma 3.5).

□

Proof of Lemma 3.7. For $p \geq 1$, application of Ito's lemma and taking expectation yields

$$\mathbb{E}\left[V_p\left(\bar{\zeta}_t^{\lambda,n}\right)\right] = \mathbb{E}\left[V_p\left(\bar{\theta}_{nT}^{\lambda}\right)\right]+\int_{nT}^{t}\mathbb{E}\left[\lambda\frac{\Delta V_p\left(\bar{\zeta}_s^{\lambda,n}\right)}{\beta}-\lambda\left\langle h\left(\bar{\zeta}_s^{\lambda,n}\right),\nabla V_p\left(\bar{\zeta}_s^{\lambda,n}\right)\right\rangle\right]\mathrm{d}s.$$

Differentiating both sides and using Lemma 3.6, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[V_p\left(\bar{\zeta}_t^{\lambda,n}\right)\right] = \mathbb{E}\left[\lambda\frac{\Delta V_p\left(\bar{\zeta}_t^{\lambda,n}\right)}{\beta}-\lambda\left\langle h\left(\bar{\zeta}_t^{\lambda,n}\right),\nabla V_p\left(\bar{\zeta}_t^{\lambda,n}\right)\right\rangle\right] \leq -\lambda\bar{c}(p)\mathbb{E}\left[V_p\left(\bar{\zeta}_t^{\lambda,n}\right)\right]+\lambda\tilde{c}(p)$$

which yields

$$\begin{aligned}\mathbb{E}\left[V_p\left(\bar{\zeta}_t^{\lambda,n}\right)\right] \leq & e^{-\lambda(t-nT)\bar{c}(p)}\mathbb{E}\left[V_p\left(\bar{\theta}_{nT}^{\lambda}\right)\right]+\frac{\tilde{c}(p)}{\bar{c}(p)}\left(1-e^{-\lambda\bar{c}(p)(t-nT)}\right) \\ \leq & e^{-\lambda(t-nT)\bar{c}(p)}\mathbb{E}\left[V_p\left(\bar{\theta}_{nT}^{\lambda}\right)\right]+\frac{\tilde{c}(p)}{\bar{c}(p)}.\end{aligned}$$

For p=2:

$$\begin{aligned}\mathbb{E}\left[V_2\left(\bar{\zeta}_t^{\lambda,n}\right)\right] \leq & e^{-\lambda(t-nT)\bar{c}(2)}\mathbb{E}\left[V_2\left(\bar{\theta}_{nT}^{\lambda}\right)\right]+\frac{\tilde{c}(2)}{\bar{c}(2)} \\ \leq & (1-\sqrt{\lambda}\frac{\eta}{2})^{nT}e^{-\lambda(t-nT)\bar{c}(2)}\mathbb{E}\left[V_2\left(\theta_0\right)\right]+\frac{\tilde{c}(2)}{\bar{c}(2)} \\ & +2\left(C_X\eta^{-1}+2M_0^2(2+\eta)+2d(\eta\beta)^{-1}\sqrt{\lambda_{max}}\right)+1 \\ \leq & \mathbb{E}\left[V_2\left(\theta_0\right)\right]+\frac{\tilde{c}(2)}{\bar{c}(2)}+2\left(C_X\eta^{-1}+2M_0^2(2+\eta)+2d(\eta\beta)^{-1}\sqrt{\lambda_{max}}\right)+1.\end{aligned}$$

For p=4:

$$\begin{aligned}\mathbb{E}\left[V_4\left(\bar{\zeta}_t^{\lambda,n}\right)\right] \leq & e^{-\lambda(t-nT)\bar{c}(4)}\mathbb{E}\left[V_4\left(\bar{\theta}_{nT}^{\lambda}\right)\right]+\frac{\tilde{c}(4)}{\bar{c}(4)} \\ \leq & 2\mathbb{E}|\theta_0|^4+2+2C`_2+\frac{\tilde{c}(4)}{\bar{c}(4)}.\end{aligned}$$

□

Proof of Proposition 3.8. See (Chau et al., 2019, Lemma 3.26.).

□

Lemma A.3. *The contraction constant in Proposition 3.8 is given by*

$$\dot{c} = \min\{\bar{\phi}, \bar{c}(p), 4\tilde{c}(p)\epsilon\bar{c}(p)\}/2$$

where the explicit expressions for $\bar{c}(p)$ and $\tilde{c}(p)$ can be found in Lemma 3.6 and $\bar{\phi}$ is given by

$$\bar{\phi} = \left(\sqrt{4\pi/K_1}\bar{b}\exp\left((\bar{b}\sqrt{K_1}/2 + 2/\sqrt{K_1})^2\right)\right)^{-1}$$

Furthermore, any ϵ can be chosen which satisfies the following inequality

$$\epsilon \leq 1 \wedge \left(8\tilde{c}(p)\sqrt{\pi/K_1}\int_0^{\tilde{b}}\exp\left((s\sqrt{K_1}/2 + 2/\sqrt{K_1})^2\right)\mathrm{d}s\right)^{-1}$$

where $K_1 = a$, $\tilde{b} = \sqrt{2\tilde{c}(p)/\bar{c}(p) - 1}$ and $\bar{b} = \sqrt{4\tilde{c}(p)(1+\bar{c}(p))/\bar{c}(p) - 1}$. The constant $\hat{c}$ is given as the ratio $\mathrm{C}_{11}/\mathrm{C}_{10}$, where C_{11}, C_{10} are given explicitly in (Chau et al., 2019, Lemma 3.26).

Proof of Lemma 3.9. One initially observes that

$$\begin{aligned}
|\bar{\theta}_t^\lambda - \bar{\zeta}_t^{\lambda,n}|^2 &= -2\lambda\int_{nT}^t \langle\bar{\zeta}_s^{\lambda,n} - \bar{\theta}_s^\lambda, h(\bar{\zeta}_s^{\lambda,n}) - H_\lambda(\theta_{\lfloor s\rfloor}, X_{\lceil s\rceil})\rangle \\
&= -2\lambda\int_{nT}^t \langle\bar{\zeta}_s^{\lambda,n} - \bar{\theta}_t^\lambda, h(\bar{\zeta}_s^{\lambda,n}) - H(\bar{\theta}_s^\lambda, X_{\lceil s\rceil})\rangle ds \\
&\quad - 2\lambda\int_{nT}^t \langle\bar{\zeta}_s^{\lambda,n} - \bar{\theta}_s^\lambda, H(\bar{\theta}_s^\lambda, X_{\lceil s\rceil}) - H(\bar{\theta}_{\lfloor s\rfloor}^\lambda, X_{\lceil s\rceil})\rangle ds \\
&\quad - 2\lambda\int_{nT}^t \langle\bar{\zeta}_s^{\lambda,n} - \bar{\theta}_s^\lambda, H(\bar{\theta}_{\lfloor s\rfloor}^\lambda, X_{\lceil s\rceil}) - H_\lambda(\theta_{\lfloor s\rfloor}, X_{\lceil s\rceil})\rangle ds.
\end{aligned}$$

Taking expectations on both sides yields that

$$\begin{aligned}
\mathbb{E}\,|\bar{\theta}_t^\lambda - \bar{\zeta}_t^{\lambda,n}|^2 &= -2\lambda\int_{nT}^t \mathbb{E}\langle\bar{\zeta}_s^{\lambda,n} - \bar{\theta}_s^\lambda, h(\bar{\zeta}_s^{\lambda,n}) - H(\bar{\theta}_s^\lambda, X_{\lceil s\rceil})\rangle ds \\
&\quad - 2\lambda\int_{nT}^t \mathbb{E}\langle\bar{\zeta}_s^{\lambda,n} - \bar{\theta}_s^\lambda, H(\bar{\theta}_s^\lambda, X_{\lceil s\rceil}) - H(\bar{\theta}_{\lfloor s\rfloor}^\lambda, X_{\lceil s\rceil})\rangle ds \\
&\quad - 2\lambda\int_{nT}^t \mathbb{E}\langle\bar{\zeta}_s^{\lambda,n} - \bar{\theta}_s^\lambda, H(\bar{\theta}_{\lfloor s\rfloor}^\lambda, X_{\lceil s\rceil}) - H_\lambda(\theta_{\lfloor s\rfloor}, X_{\lceil s\rceil})\rangle ds \\
&= \int_{nT}^t A_s + B_s + D_s ds
\end{aligned} \tag{45}$$

Using the property in Proposition 2.6, one obtains

$$\begin{aligned}
A_t &= -2\lambda\,\mathbb{E}\left[\mathbb{E}\langle\bar{\zeta}_t^{\lambda,n} - \bar{\theta}_t^\lambda, h(\bar{\zeta}_t^{\lambda,n}) - H(\bar{\theta}_t^\lambda, X_{\lceil t\rceil})\rangle\,\middle|\,\bar{\zeta}_t^{\lambda,n}, \bar{\theta}_t^\lambda\right] \\
&= -2\lambda\,\mathbb{E}\langle\bar{\zeta}_t^{\lambda,n} - \bar{\theta}_t^\lambda, h(\bar{\zeta}_t^{\lambda,n}) - h(\bar{\theta}_t^\lambda)\rangle \\
&\leq 2\lambda a\,\mathbb{E}\,|\bar{\zeta}_t^{\lambda,n} - \bar{\theta}_t^\lambda|^2.
\end{aligned} \tag{46}$$

In addition, taking advantage of the polynomial Lipschitzness of H, one observes that

$$\begin{aligned}
B_t &\leq \frac{\lambda a}{2}\,\mathbb{E}\,|\bar{\zeta}_t^{\lambda,n} - \bar{\theta}_t^\lambda|^2 + \frac{2L\lambda}{a}\,\mathbb{E}\left[(1+|X_{\lceil t\rceil}|)^{2\rho}(1+|\bar{\theta}_t^\lambda| + |\bar{\theta}_{\lfloor t\rfloor}^\lambda|)^{2l}|\bar{\theta}_t^\lambda - \bar{\theta}_{\lfloor t\rfloor}^\lambda|^2\right] \\
&\leq \frac{\lambda a}{2}\,\mathbb{E}\,|\bar{\zeta}_t^{\lambda,n} - \bar{\theta}_t^\lambda|^2 + \frac{2L\lambda}{a}\sqrt{\mathbb{E}\left[(1+|X_{\lceil t\rceil}|)^{4\rho}(1+|\bar{\theta}_t^\lambda| + |\bar{\theta}_{\lfloor t\rfloor}^\lambda|)^{4l}\right]}\sqrt{\mathbb{E}\left[|\bar{\theta}_t^\lambda - \bar{\theta}_{\lfloor t\rfloor}^\lambda|^4\right]} \\
&\leq \frac{\lambda a}{2}\,\mathbb{E}\,|\bar{\zeta}_t^{\lambda,n} - \bar{\theta}_t^\lambda|^2 + \frac{2L\lambda}{a}\sqrt{\mathbb{E}\left[(1+|X_{\lceil t\rceil}|)^{4\rho}(1+2|\bar{\theta}_{\lfloor t\rfloor}^\lambda| + |\bar{\theta}_t^\lambda - \bar{\theta}_{\lfloor t\rfloor}^\lambda|)^{4l}\right]}\sqrt{\mathbb{E}\left[|\bar{\theta}_t^\lambda - \bar{\theta}_{\lfloor t\rfloor}^\lambda|^4\right]}
\end{aligned}$$

Furthermore, one applies again the Cauchy–Schwarz inequality to obtain

$$B_t \leq \frac{\lambda a}{2}\,\mathbb{E}\,|\bar{\zeta}_t^{\lambda,n} - \bar{\theta}_t^\lambda|^2 + \frac{2L}{a}(\mathbb{E}(1+|X_0|)^{8\rho})^{\frac{1}{4}}9^l\left(1 + 2^{8l}\,\mathbb{E}\,|\bar{\theta}_{\lfloor t\rfloor}^\lambda|^{8l} + \mathbb{E}\,|\bar{\theta}_t^\lambda - \bar{\theta}_{\lfloor t\rfloor}^\lambda|^{8l}\right)^{\frac{1}{4}}\lambda\sqrt{\mathbb{E}\,|\bar{\theta}_t^\lambda - \bar{\theta}_{\lfloor t\rfloor}^\lambda|^4}. \tag{47}$$

By taking into consideration that

$$\begin{aligned}
|\bar{\theta}_t^\lambda - \bar{\theta}_{\lfloor t\rfloor}^\lambda| &\leq \lambda\left|\int_{\lfloor t\rfloor}^t H_\lambda(\bar{\theta}_{\lfloor u\rfloor}^\lambda, X_{\lceil u\rceil})du\right| + \sqrt{\frac{2\lambda}{\beta}}|\tilde{B}_t^\lambda - \tilde{B}_{\lfloor t\rfloor}^\lambda| \\
&\leq \sqrt{\lambda}\left(\int_{\lfloor t\rfloor}^t K(X_{\lceil u\rceil}) + \eta|\bar{\theta}_{\lfloor u\rfloor}^\lambda|du + \sqrt{\frac{2}{\beta}}|\tilde{B}_t^\lambda - \tilde{B}_{\lfloor t\rfloor}^\lambda|\right),
\end{aligned}$$

and that both the requires moments of $X_{\lceil t\rceil}$ and of $\bar{\theta}_{\lfloor t\rfloor}^\lambda$ are finite due to Assumption 2 and (13) respectively, one deduces that $\sqrt{\mathbb{E}\,|\bar{\theta}_t^\lambda - \bar{\theta}_{\lfloor t\rfloor}^\lambda|^4} \leq \tilde{C}_1\lambda$, where $\tilde{C}_1 = 9\sqrt{\mathbb{E}[K^4(X_0)] + \eta^4(\mathbb{E}\,|\theta_0|^4 + C\text{`}_2) + \frac{48}{\beta^4}d^2}$. Similarly, $\mathbb{E}\,|\bar{\theta}_t^\lambda - \bar{\theta}_{\lfloor t\rfloor}^\lambda|^{8l} \leq$

$\tilde{C}_2\lambda^{4l}$, where $\tilde{C}_2 = 3^{4l}\sqrt{\mathbb{E}\,K^{8l}(X_0) + \eta^{8l}(\mathbb{E}\,|\theta_0|^{8l} + C\text{`}_{4l} + (\frac{2}{\beta})^{8l}d^{4l}(8l-1)!!}$. Here, the fact was used that the increment of a d-dimensional Brownian motion has a d-dimensional Gaussian distribution with mean 0 and covariance matrix $(t - \lfloor t \rfloor)\mathbb{I}_d$. Its $2m$-th moment is given by

$$\mathbb{E}\left[|\tilde{B}_t^\lambda - \tilde{B}_{\lfloor t \rfloor}^\lambda|^{2m}\right] = \mathbb{E}\left[\left(\sum_{i=1}^d Y_i^2\right)^m\right] \leq d^m\,\mathbb{E}\,Z^{2m},$$

where Y_i, $i \in \{1, \ldots, d\}$, are the increments of the one dimensional Brownian motions which follow the same distribution as $Z \sim \mathcal{N}(0, t - \lfloor t \rfloor)$. Hence,

$$\mathbb{E}\left[|\tilde{B}_t^\lambda - \tilde{B}_{\lfloor t \rfloor}^\lambda|^{2m}\right] \leq d^m(2m-1)!!(t - \lfloor t \rfloor)^m \leq d^m(2m-1)!!.$$

Thus, (47) implies that

$$B_t \leq \frac{\lambda a}{2}\,\mathbb{E}\,|\bar{\zeta}_t^{\lambda,n} - \bar{\theta}_t^\lambda|^2 + C_1\lambda^2 \tag{48}$$

where

$$C_1 = 2\frac{L}{a}2^{4l}\tilde{C}_1(\mathbb{E}(1+X_0)^{8\rho})^{\frac{1}{4}}\left(1 + 2^{8l}(\mathbb{E}\,|\theta_0|^{8l} + C\text{`}_{4l}) + \tilde{C}_2\right)^{\frac{1}{4}}. \tag{49}$$

Moreover,

$$\begin{aligned}
D_t \leq& \frac{\lambda a}{2}\,\mathbb{E}\,|\bar{\theta}_t^\lambda - \bar{\zeta}_t^{\lambda,n}|^2 + \frac{2\lambda}{a}\,\mathbb{E}\,|H(\bar{\theta}_{\lfloor t \rfloor}^\lambda, X_{\lceil t \rceil}) - H_\lambda(\theta_{\lfloor t \rfloor}, X_{\lceil t \rceil})|^2 \\
\leq& \frac{\lambda a}{2}\,\mathbb{E}\,|\bar{\theta}_t^\lambda - \bar{\zeta}_t^{\lambda,n}|^2 + \frac{2\lambda^2}{a}\,\mathbb{E}\left[|H(\bar{\theta}_{\lfloor t \rfloor}^\lambda, X_{\lceil t \rceil})||\bar{\theta}_{\lfloor t \rfloor}^\lambda||^{2r}\right]^2 \\
\leq& \frac{\lambda a}{2}\,\mathbb{E}\,|\bar{\theta}_t^\lambda - \bar{\zeta}_t^{\lambda,n}|^2 + \frac{4\lambda^2}{a}\left(\mathbb{E}\,|H(\bar{\theta}_{\lfloor t \rfloor}^\lambda, X_{\lceil t \rceil}) - H(\bar{\theta}_0^\lambda, X_{\lceil t \rceil})|^2|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^{4r} + \mathbb{E}\,|H(\bar{\theta}_0^\lambda, X_{\lceil t \rceil})|^2|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^{4r}\right) \\
\leq& \frac{\lambda a}{2}\,\mathbb{E}\,|\bar{\theta}_t^\lambda - \bar{\zeta}_t^{\lambda,n}|^2 + \frac{L^2\lambda^2}{a}\,\mathbb{E}\left[(1+|X_{\lceil t \rceil}|)^{2\rho}(1+|\theta_0| + |\bar{\theta}_{\lfloor t \rfloor}^\lambda|)^{2l}|\theta_0 - \bar{\theta}_{\lfloor t \rfloor}^\lambda|^2|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^{4r}\right] \\
&+ \frac{4\lambda^2}{a}\,\mathbb{E}\,|H(\theta_0, X_0)|^2|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^{4r} \\
\leq& \frac{\lambda a}{2}\,\mathbb{E}\,|\bar{\theta}_t^\lambda - \bar{\zeta}_t^{\lambda,n}|^2 + \frac{L^2\lambda^2}{a}\sqrt{\mathbb{E}\left[(1+|X_{\lceil t \rceil}|)^{4\rho}(1+|\theta_0| + |\bar{\theta}_{\lfloor t \rfloor}^\lambda|)^{4l}|\theta_0 - \bar{\theta}_{\lfloor t \rfloor}^\lambda|^4\right]}\sqrt{\mathbb{E}\,|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^{8r}} \\
&+ \frac{4\lambda^2}{a}\sqrt{\mathbb{E}\,|H(\theta_0, X_0)|^4}\sqrt{\mathbb{E}\,|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^{8r}} \\
\leq& \frac{\lambda a}{2}\,\mathbb{E}\,|\bar{\theta}_t^\lambda - \bar{\zeta}_t^{\lambda,n}|^2 + \sqrt{C\text{`}_{4r} + \mathbb{E}\,|\theta_0|^{8r}} \\
&\times \left(\frac{L^2\lambda^2}{a}(\mathbb{E}(1+|X_0|)^{8\rho})^{\frac{1}{4}}\left(\mathbb{E}(1+|\theta_0| + \bar{\theta}_{\lfloor t \rfloor}^\lambda)^{8l}|\theta_0 - \bar{\theta}_{\lfloor t \rfloor}^\lambda|^8\right)^{\frac{1}{4}}\right) \\
&+ \sqrt{C\text{`}_{4r} + \mathbb{E}\,|\theta_0|^{8r}}\frac{4\lambda^2}{a}\sqrt{\mathbb{E}\,|H(\theta_0, X_0)|^4} \\
\leq& \frac{\lambda a}{2}\,\mathbb{E}\,|\bar{\theta}_t^\lambda - \bar{\zeta}_t^{\lambda,n}|^2 + \sqrt{C\text{`}_{4r} + \mathbb{E}\,|\theta_0|^{8r}} \\
&\times \frac{L^2\lambda^2}{a}(\mathbb{E}(1+|X_0|)^{8\rho})^{\frac{1}{4}}2^{2l+2}\left(\mathbb{E}(1+|\theta_0|)^{16l} + \mathbb{E}\,|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^{16l}\right)^{\frac{1}{8}}\left(\mathbb{E}\,|\theta_0|^{16} + \mathbb{E}\,|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^{16}\right)^{\frac{1}{8}} \\
&+ \sqrt{C\text{`}_{4r} + \mathbb{E}\,|\theta_0|^{8r}}\frac{4\lambda^2}{a}\sqrt{\mathbb{E}\,|H(\theta_0, X_0)|^4} \\
\leq& \frac{\lambda a}{2}\,\mathbb{E}\,|\bar{\theta}_t^\lambda - \bar{\zeta}_t^{\lambda,n}|^2 + C_2\lambda^2
\end{aligned} \tag{50}$$

where

$$\begin{aligned}
C_2 =& \sqrt{C\text{`}_{4r} + \mathbb{E}\,|\theta_0|^{8r}}\left(\frac{L^2}{a}(\mathbb{E}(1+|X_0|)^{8\rho})^{\frac{1}{4}}2^{2l+2}\left(\mathbb{E}(1+|\theta_0|)^{16l} + \mathbb{E}\,|\theta_0|^{16l} + C\text{`}_{8l}\right)^{\frac{1}{8}}\right. \\
&\left. + \left(\mathbb{E}\,|\theta_0|^{16} + \mathbb{E}\,|\theta_0|^{16} + C\text{`}_8\right)^{\frac{1}{8}}\right) + \frac{4}{a}\sqrt{\mathbb{E}\,|H(\theta_0, X_0)|^4}.
\end{aligned} \tag{51}$$

In view of the estimates (46), (48) and (50), one concludes that equation (45) can be rewritten as

$$\mathbb{E}\,|\bar{\theta}_t^\lambda - \bar{\zeta}_t^{\lambda,n}|^2 \leq 3\lambda a\int_{nT}^t \mathbb{E}\,|\bar{\theta}_s^\lambda - \bar{\zeta}_s^{\lambda,n}|^2 ds + (C_1 + C_2)\lambda < \infty.$$

The application of Gronwall's Lemma implies that

$$\mathbb{E}\,|\bar{\theta}_t^\lambda - \bar{\zeta}_t^{\lambda,n}|^2 \leq c\lambda, \qquad \text{where } c = e^{3a}(C_1 + C_2)$$

which yields the desired rate while the constant c is independent of t and λ. □

Proof of Lemma 3.10. In view of the result in Lemma 3.9,

$$
\begin{aligned}
&W_1\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,n}\right),\mathcal{L}\left(Z_t^{\lambda}\right)\right)\\
&\leq \sum_{k=1}^{n} W_1\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,k}\right),\mathcal{L}\left(\bar{\zeta}_t^{\lambda,k-1}\right)\right)\\
&\leq \sum_{k=1}^{n} w_{1,2}\left(\mathcal{L}\left(\zeta_t^{kT,\bar{\theta}^{\lambda}_{kT},\lambda}\right),\mathcal{L}\left(\zeta_t^{kT,\bar{z}^{\lambda,k-1}_{kT},\lambda}\right)\right)\\
&\leq \hat{c}\sum_{k=1}^{n}\exp(-\dot{c}(n-k))w_{1,2}\left(\mathcal{L}\left(\bar{\theta}^{\lambda}_{kT}\right),\mathcal{L}\left(\bar{\zeta}^{\lambda,k-1}_{kT}\right)\right)\\
&\leq \hat{c}\sum_{k=1}^{n}\exp(-\dot{c}(n-k))W_2\left(\mathcal{L}\left(\bar{\theta}^{\lambda}_{kT}\right),\mathcal{L}\left(\bar{\zeta}^{\lambda,k-1}_{kT}\right)\right)\left[1+\left\{\mathbb{E}\left[V_4\left(\bar{\theta}^{\lambda}_{kT}\right)\right]\right\}^{1/2}+\left\{\mathbb{E}\left[V_4\left(\bar{\zeta}^{\lambda,k-1}_{kT}\right)\right]\right\}^{1/2}\right]\\
&\leq (\sqrt{\lambda})^{-1}\hat{c}\sum_{k=1}^{n}\exp(-\dot{c}(n-k))W_2^2\left(\mathcal{L}\left(\bar{\theta}^{\lambda}_{kT}\right),\mathcal{L}\left(\bar{\zeta}^{\lambda,k-1}_{kT}\right)\right)\\
&+3\sqrt{\lambda}\hat{c}\sum^{n}\exp(-\dot{c}(n-k))\left[1+\mathbb{E}\left[V_4\left(\bar{\theta}^{\lambda}_{kT}\right)\right]+\mathbb{E}\left[V_4\left(\bar{\zeta}^{\lambda,k-1}_{kT}\right)\right]\right]\\
&\leq \sqrt{e^{3a}(C_1+C_2)}\sqrt{\lambda}\frac{\hat{c}}{1-\exp(-\dot{c})}\\
&+3\sqrt{\lambda}\frac{\hat{c}}{1-\exp(-\dot{c})}\left(1+2\mathbb{E}|\theta_0|^4+2+2C`_2+\frac{\tilde{c}(4)}{\bar{c}(4)}+2\mathbb{E}|\theta_0|^4+2+2C`_2\right)\\
&=\sqrt{\lambda}z_1
\end{aligned}
$$

where

$$
z_1=\frac{\hat{c}}{1-exp(-\dot{c})}\left[\sqrt{e^{3a}(C_1+C_2)}+3\left(5+4C`_2\frac{\tilde{c}(4)}{\bar{c}(4)}+4\mathbb{E}|\theta_0|^4\right)\right] \tag{52}
$$

and C_1, C_2 are given by (49) , (51) respectively. □

A.3 Proof of main results

Lemma A.4. *Let Assumptions 1 and 2 hold. Then for* $0<\lambda\leq\lambda_{\max}$, $t\in[nT,(n+1)T]$,

$$
W_1\left(\mathcal{L}\left(\bar{\theta}_t^{\lambda}\right),\mathcal{L}\left(Z_t^{\lambda}\right)\right)\leq\sqrt{\lambda}\sqrt{e^{3a}(C_1+C_2)}+\sqrt{\lambda}z_1=\sqrt{\lambda}(z_1+\sqrt{e^{3a}(C_1+C_2)})
$$

where C_1, C_2 *and* z_1 *are given by (49) , (51) and (52) respectively.*

Proof. Combining the results stated in Lemmas 3.9 and 3.10

$$
\begin{aligned}
W_1\left(\mathcal{L}\left(\bar{\theta}_t^{\lambda}\right),\mathcal{L}\left(Z_t^{\lambda}\right)\right)&\leq W_1\left(\mathcal{L}\left(\bar{\theta}_t^{\lambda}\right),\mathcal{L}\left(\bar{\zeta}_t^{\lambda,n}\right)\right)+W_1\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,n}\right),\mathcal{L}\left(Z_t^{\lambda}\right)\right)\\
&\leq W_2\left(\mathcal{L}\left(\bar{\theta}_t^{\lambda}\right),\mathcal{L}\left(\bar{\zeta}_t^{\lambda,n}\right)\right)+W_1\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,n}\right),\mathcal{L}\left(Z_t^{\lambda}\right)\right)\\
&\leq\sqrt{\lambda}\sqrt{e^{3a}(C_1+C_2)}+\sqrt{\lambda}z_1\\
&=\sqrt{\lambda}(z_1+\sqrt{e^{3a}(C_1+C_2)}),
\end{aligned}
$$

which yields the desired result. □

Proof of Theorem 2.9. By taking into consideration the result in the Lemma A.4 and the property of $w_{1,2}$ in Proposition 3.8, one calculates

$$
\begin{aligned}
W_1\left(\mathcal{L}\left(\theta_t^{\lambda}\right),\pi_\beta\right)&\leq W_1\left(\mathcal{L}\left(\bar{\theta}_t^{\lambda}\right),\mathcal{L}\left(Z_t^{\lambda}\right)\right)+W_1\left(\mathcal{L}\left(Z_t^{\lambda}\right),\pi_\beta\right)\\
&\leq\sqrt{\lambda}(z_1+\sqrt{e^{3a}(C_1+C_2)})+\hat{c}e^{-\dot{c}\lambda t}w_{1,2}\left(\theta_0,\pi_\beta\right)\\
&\leq\sqrt{\lambda}(z_1+\sqrt{e^{3a}(C_1+C_2)})+\hat{c}e^{-\dot{c}\lambda t}\left[1+\mathbb{E}\left[V_2\left(\theta_0\right)\right]+\int_{\mathbb{R}^d}V_2(\theta)\pi_\beta(d\theta)\right]\\
&\leq\sqrt{\lambda}(z_1+\sqrt{e^{3a}(C_1+C_2)})+\hat{c}e^{-\dot{c}n}\left[1+\mathbb{E}\left[V_2\left(\theta_0\right)\right]+\int_{\mathbb{R}^d}V_2(\theta)\pi_\beta(d\theta)\right]
\end{aligned}
$$

where C_1, C_2 and z_1 are given by (49) , (51) and (52) respectively. □

Lemma A.5. *Let Assumptions 1 and 2 hold. Then for* $0<\lambda\leq\lambda_{max}$,$t\in[nT,(n+1)T]$ *there holds*

$$
W_2\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,n}\right),\mathcal{L}\left(Z_t^{\lambda}\right)\right)\leq\lambda^{\frac{1}{4}}z_2
$$

where z_2 *is given by (53).*

Proof. Using that $W_2 \leq \sqrt{2w_{1,2}}$, one obtains

$$\begin{aligned}
W_2\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,n}\right),\mathcal{L}\left(Z_t^{\lambda}\right)\right) \leq & \sum_{k=1}^{n} W_2\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,k}\right),\mathcal{L}\left(\bar{\zeta}_t^{\lambda,k-1}\right)\right) \\
\leq & \sum_{k=1}^{n} \sqrt{2} w_{1,2}^{1/2}\left(\mathcal{L}\left(\zeta_t^{kT,\bar{\theta}_{kT}^{\lambda},\lambda}\right),\mathcal{L}\left(\zeta_t^{kT,\bar{\zeta}_{kT}^{\lambda,k-1},\lambda}\right)\right) \\
\leq & \sqrt{2\dot{c}} \sum_{k=1}^{n} \exp(-\dot{c}(n-k)/2) W_2^{1/2}\left(\mathcal{L}\left(\bar{\theta}_{kT}^{\lambda}\right),\mathcal{L}\left(\bar{\zeta}_{kT}^{\lambda,k-1}\right)\right) \\
& \times\left[1+\left\{\mathbb{E}\left[V_4\left(\bar{\theta}_{kT}^{\lambda}\right)\right]\right\}^{1/2}+\left\{\mathbb{E}\left[V_4\left(\bar{\zeta}_{kT}^{\lambda,k-1}\right)\right]\right\}^{1/2}\right]^{1/2} \\
\leq & \lambda^{-1/4}\sqrt{2\dot{c}} \sum_{k=1} \exp(-\dot{c}(n-k)/2) W_2\left(\mathcal{L}\left(\bar{\theta}_{kT}^{\lambda}\right),\mathcal{L}\left(\bar{\zeta}_{kT}^{\lambda,k-1}\right)\right) \\
& + \lambda^{1/4}\sqrt{2\dot{c}} \sum_{k=1}^{n} \exp(-\dot{c}(n-k)/2) \\
& \times\left[1+\left\{\mathbb{E}\left[V_4\left(\bar{\theta}_{kT}^{\lambda}\right)\right]\right\}^{1/2}+\left\{\mathbb{E}\left[V_4\left(\bar{\zeta}_{kT}^{\lambda,k-1}\right)\right]\right\}^{1/2}\right] \\
= & \lambda^{\frac{1}{4}}\sqrt{2\dot{c}}\frac{1}{1-exp(-\dot{c}/2)} e^{3a}(C_1+C_2) \\
& + \lambda^{\frac{1}{4}}\sqrt{2\dot{c}}\frac{1}{1-exp(-\dot{c}/2)} \\
& \times\left[1+\sqrt{2\mathbb{E}|\theta_0|^4+2+2C`_2+\frac{\tilde{c}(4)}{\bar{c}(4)}}+\sqrt{2\mathbb{E}|\theta_0|^4+2+2C`_2}\right] \\
= & \lambda^{\frac{1}{4}} z_2
\end{aligned}$$

where

$$z_2 = \sqrt{2\dot{c}}\frac{1}{1-exp(-\dot{c}/2)}\left[e^{3a}(C_1+C_2)+1+\sqrt{2\mathbb{E}|\theta_0|^4+2+2C`_2+\frac{\tilde{c}(4)}{\bar{c}(4)}}+\sqrt{2\mathbb{E}|\theta_0|^4+2+2C`_2}\right]. \tag{53}$$

□

Proof of Corollary 2.10. Combining Lemma 3.9 and Lemma A.5, one obtains

$$\begin{aligned}
W_2\left(\mathcal{L}\left(\theta_t^{\lambda}\right),\pi_\beta\right) \leq & W_2\left(\mathcal{L}\left(\theta_t^{\lambda}\right),\mathcal{L}\left(Z_t^{\lambda}\right)\right)+W_2\left(\mathcal{L}\left(Z_t^{\lambda}\right),\pi_\beta\right) \\
\leq & W_2\left(\mathcal{L}\left(\bar{\theta}_t^{\lambda}\right),\mathcal{L}\left(\bar{\zeta}_t^{\lambda,n}\right)\right)+W_2\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,n}\right),\mathcal{L}\left(Z_t^{\lambda}\right)\right)+W_2\left(\mathcal{L}\left(Z_t^{\lambda}\right),\pi_\beta\right) \\
\leq & \sqrt{e^{3a}(C_1+C_2)}\sqrt{\lambda}+z_2\lambda^{\frac{1}{4}}+\sqrt{2w_{1,2}\left(\mathcal{L}\left(Z_t^{\lambda}\right),\pi_\beta\right)} \\
\leq & \sqrt{e^{3a}(C_1+C_2)}\sqrt{\lambda}+z_2\lambda^{\frac{1}{4}}+\hat{c}^{1/2}e^{-\dot{c}\lambda t/2}\sqrt{2w_{1,2}\left(\theta_0,\pi_\beta\right)} \\
\leq & \sqrt{e^{3a}(C_1+C_2)}\sqrt{\lambda}+z_2\lambda^{\frac{1}{4}}+\sqrt{2}\hat{c}^{1/2}e^{-\dot{c}\lambda t/2} \\
& \times\left(1+\mathbb{E}\left[V_2\left(\theta_0\right)\right]+\int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(d\theta)\right)^{1/2} \\
\leq & \sqrt{e^{3a}(C_1+C_2)}\sqrt{\lambda}+z_2\lambda^{\frac{1}{4}}+\sqrt{2}\hat{c}^{1/2}e^{-\dot{c}n/2}\left(1+\mathbb{E}\left[V_2\left(\theta_0\right)\right]+\int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(d\theta)\right)^{1/2}.
\end{aligned}$$

□

Proof of Lemma 3.1. Taking into account that

$$h(\theta) = \mathbb{E}[G(\theta, X_0)] + \eta|\theta|^{2r}\theta$$

and the polynomial growth of G in (2), there exist $r_1 = \mathbb{E}[K(X_0)] + \eta$, $r_2 = 2\,\mathbb{E}[K(X_0)]$, such that

$$|h(\theta)| \leq r_1|\theta|^l + r_2 \quad \forall\theta \in \mathbb{R}^d,$$

where $l = 2r + 1$. As a result,

$$\begin{aligned} u(w) - u(v) &= \int_0^1 \langle w - v, \nabla u((1-t)v + tw)\rangle \mathrm{d}t \\ &\leq \int_0^1 |\nabla u((1-t)v + tw)||w - v| \mathrm{d}t \\ &\leq \int_0^1 \left(a_1(1-t)^l |v|^l + a_1 t^l |w|^l + r_2\right) |w - v| \mathrm{d}t \\ &= \left(\frac{a_1}{l+1}|v|^l + \frac{a_1}{l+1}|w|^l + r_2\right) |w - v| \end{aligned}$$

where $a_1 = 2^l r_1$. Let P the coupling of μ and ν that achieves $W_2(\mu, \nu)$, that is $P = (\mathcal{L}(W), \mathcal{L}(V))$ with $\mu = \mathcal{L}(W)$ and $\nu = \mathcal{L}(V)$. Taking a closer look one notices that

$$\begin{aligned} \int_{\mathbb{R}^d} u \mathrm{d}\mu - \int_{\mathbb{R}^d} u \mathrm{d}v &= \mathbb{E}_P[u(W) - u(V)] \\ &\leq \sqrt{\mathbb{E}_P\left(\frac{a_1}{l+1}|W|^l + \frac{a_1}{l+1}|V|^l + r_2\right)^2} \cdot \sqrt{\mathbb{E}_{\mathrm{P}}\left[|W - V|^2\right]} \\ &\leq \left(\frac{a_1}{l+1}\sqrt{\mathbb{E}\,|W|^{2l}} + \frac{a_1}{l+1}\sqrt{\mathbb{E}\,|V|^{2l}} + r_2\right) \cdot \mathcal{W}_2(\mu, v) \end{aligned}$$

Applying this to the particular case where $W = \theta_n^\lambda$ and $V = \theta_\infty$ yields

$$\mathbb{E}\,u(\theta_n^\lambda) - \mathbb{E}\,u(\theta_\infty) \leq \left(\frac{a_1}{l+1}\sqrt{\mathbb{E}\,|\theta_0|^{2l} + C`_l} + \frac{a_1}{l+1}\sqrt{\sigma_{2l}} + r_2\right) W_2\left(\mathcal{L}\left(\theta_n^\lambda\right), \pi_\beta\right)$$

where σ_{2l} is the $2l$-moment of π_β. □

Proof of Lemma 3.2. A similar approach as in (Raginsky et al., 2017, Section 3.5) is employed here, however due to the difference in the smoothness condition for H (and consequently for h), see our Proposition 2.7 in contrast to global Lipschitzness which is required in Raginsky et al. (2017), we provide the details for obtaining a bound for $\log \Lambda$. Recall that Λ represents the normalizing constant, i.e.

$$\Lambda := \int_{\mathbb{R}^d} e^{-\beta u(\theta)} d\theta.$$

Initially, one observes that due to the monotonicity condition (3),

$$\langle \theta^*, h(\theta^*)\rangle \geq A|\theta^*|^2 - B \implies |\theta^*| \leq \sqrt{\frac{B}{A}} \leq R_0.$$

Consequently, one calculates that

$$\begin{aligned} u_* - u(w) &= \int_0^1 \langle h(w + t(\theta^* - w), \theta^* - w\rangle dt \\ &= \int_0^1 \langle h(w + t(\theta^* - w) - h(\theta^*), \theta^* - w\rangle dt \\ &= \int_0^1 \frac{1}{t-1}\langle h(w + t(\theta^* - w) - h(\theta^*), w - \theta^* + t(\theta^* - w)\rangle dt, \end{aligned}$$

which due to the polynomial lipschitzness of h yields that

$$\begin{aligned} -\beta(u_* - u(w)) &= \beta|u_* - u(w)| \\ &\leq \beta \int_0^1 \frac{1}{1-t}\left|\langle h\left(w + t(\theta^* - w)\right) - h(\theta^*), w - \theta^* + t(\theta^* - w)\rangle\right| dt \\ &\leq \int_0^1 b'(1 + |w| + |\theta^* - w| + |\theta^*|)^l (1-t)|w - \theta^*|^2 dt \\ &= b'(1 + 2|\theta^*| + 2|\theta^* - w|)^l \frac{|w - \theta^*|^2}{2}, \end{aligned} \tag{54}$$

where $b' = L\mathbb{E}(1 + |X_0|)^\rho \beta$. As a result,

$$\begin{aligned} I = \int_{\mathbb{R}^d} e^{\beta(u_* - u(w))} dw &\geq \int_{\mathbb{R}^d} e^{-b'(1+2|w-\theta^*|+2|\theta^*|)^l (\frac{|w-\theta^*|^2}{2})} dw \\ &\geq \int_{\bar{B}(\theta^*, R_0)} e^{-b'(1+4R_0)^l (\frac{|w-\theta^*|^2}{2})} dw \\ &= \left(\frac{2\pi}{b''}\right)^{\frac{d}{2}} \int_{\bar{B}(\theta^*, R_0)} f_X(w) dw \end{aligned}$$

where $b'' = b'(1+4R_0)^l$, f is a density function of a multivariate normal variable X with mean θ^* and covariance matrix $V = 1/b'' I_d$, where I_d is the d-dimensional identity matrix. Applying (multivariate) Chebyshev's inequality, yields

$$P\left(||X-\theta^*|| > R_0\right) = P\left(\sqrt{(X-\theta^*)^T V^{-1}(X-\theta^*)} > R_0\sqrt{b''}\right) \leq \frac{d}{R_0^2 b''}$$

which leads to

$$I \geq \left(\frac{2\pi}{b''}\right)^{\frac{d}{2}} \left(1 - \frac{d}{R_0^2 b''}\right).$$

Consequently, following (Raginsky et al., 2017, Section 3.5), one obtains

$$\log \Lambda \geq -\beta u_* + \frac{d}{2}\log\left(\frac{2\pi}{b''}\right) + \log\left(1 - \frac{d}{R_0^2 b''}\right).$$

Thus, by letting $M := b''/\beta$ and in view of (3) and (Raginsky et al., 2017, Lemma 3), one obtains

$$\mathbb{E}\, u(\theta_\infty) - u_* \leq \frac{d}{2\beta}\log\left(\frac{eM}{A}\left(\frac{B\beta}{d}+1\right)\right) - \frac{1}{\beta}\log\left(1 - \frac{d}{M\beta R_0^2}\right).$$

□

A.4 Complementary details to Section 4

We start with an easy observation about the equivalence of the operator norm and Euclidean norm of a linear operator. For any $k, l \in \mathbb{N}_+$, $W \in \mathrm{Lin}\left(\mathbb{R}^k, \mathbb{R}^l\right)$ and $z \in \mathbb{R}^k$,

$$\begin{aligned}|Wz|^2 &= \sum_{i=1}^{l}[Wz]_i^2 = \sum_{i=1}^{l}\left(\sum_{j=1}^{k} W_{ij} z_j\right)^2 \\ &\leq \left(\sum_{j=1}^{k} z_j^2\right)\sum_{i=1}^{l}\sum_{j=1}^{k} W_{ij}^2 = |z|^2 |W|^2.\end{aligned}$$

On the other hand, if $l \leq k$, then $|W|^2 = \sum_{i=1}^{l}\sum_{j=1}^{k} W_{ij}^2 = \sum_{i=1}^{l}[WW^*]_{ii} \leq l\|W\|^2$ and similarly, for $k \leq l$, $|W|^2 \leq k\|W^*\|^2 = k\|W\|^2$. As a result, we obtain

$$\|W\| \leq |W| \leq \min(\sqrt{k}, \sqrt{l})\|W\|. \tag{55}$$

In particular, if $k = 1$ or $l = 1$ then the Euclidean and operator norms coincide. As easily seen, for any $\eta \in C_b(\mathbb{R})$, $W \in \mathrm{Lin}\left(\mathbb{R}^k, \mathbb{R}^l\right)$ and $z \in \mathbb{R}^k$,

$$|\eta_W(z)| \leq \sqrt{l}\,\|\eta\|_\infty, \tag{56}$$

$$\|M_{\eta_W(z)}\| \leq \|\eta\|_\infty. \tag{57}$$

The next lemma establishes upper bound on the norm of $\partial_\theta f(\theta, \mathbf{z})$ involving an order n polynomial of $|\theta|$.

Lemma A.6. *Let $\theta = (\phi, \mathbf{w}) \in \Theta$ and $x = (\mathbf{z}, y) \in \mathbb{R}^{m-1} \times \mathbb{R}$ arbitrary. Then, for the Euclidean norm of the partial derivatives of the regression function with respect to the learning parameter, we have*

$$|\partial_\theta f(\theta, \mathbf{z})| \leq D^{1/2}\sqrt{n+1}(1+|x|)(1+|||\sigma|||)^{n+1}(1+|\theta|^n). \tag{58}$$

Furthermore, for the operator norm of the partial derivatives of nonlinear maps appearing in the definition of f, see (16), one obtains that

$$\left\|\partial_{W_i}\sigma\left(\mathbf{w}_1^n, \mathbf{z}\right)\right\| \leq \sqrt{D}(1+|x|)(1+|||\sigma|||)^{n-i+2}|\theta|^{n-i} \quad i = 1, \ldots, n \tag{59}$$

holds.

Proof. In what follows, we calculate $\partial_\theta f(\theta, \mathbf{z}) \in \Theta^*$ at a fixed $\theta \in \Theta$ and $\mathbf{z} \in \mathbb{R}^{m-1}$. For any $\tilde{\theta} = (\tilde{\phi}, \tilde{\mathbf{w}})$, where $\tilde{\mathbf{w}} = (\tilde{W}_1, \ldots, \tilde{W}_n)$,

$$\partial_\theta f(\theta, \mathbf{z})(\tilde{\theta}) = \partial_\phi f(\theta, \mathbf{z})(\tilde{\phi}) + \sum_{i=1}^{n} \partial_{W_i} f(\theta, \mathbf{z})(\tilde{W}_i).$$

The map $\phi \mapsto f((\phi, \mathbf{w}), \mathbf{z})$ is linear hence, by (56),

$$|\partial_\phi f(\theta, \mathbf{z})| = |\sigma\left(\mathbf{w}_1^n, \mathbf{z}\right)| \leq \sqrt{d_n}\,\|\sigma\|_\infty.$$

Moreover,

$$\partial_{W_i} f(\theta, \mathbf{z})(\tilde{W}_i) = \phi \circ \partial_{W_i}\sigma\left(\mathbf{w}_1^n, \mathbf{z}\right)(\tilde{W}_i).$$

Thus, by the chain rule, for $i = 1, \ldots, n$, one deduces that

$$\begin{aligned}\partial_{W_i}\sigma\left(\mathbf{w}_1^n, \mathbf{z}\right)(\tilde{W}_i) &= \left[\prod_{j=1}^{n-i} \partial_{\mathbf{z}}\sigma_{W_{n-j+1}}\left(\sigma\left(\mathbf{w}_1^{n-j}, \mathbf{z}\right)\right)\right]\partial_{W_i}\sigma_{W_i}\left(\sigma\left(\mathbf{w}_1^{i-1}, \mathbf{z}\right)\right)(\tilde{W}_i) \\ &= \left[\prod_{j=1}^{n-i} M_{\sigma'_{W_{n-j+1}}\left(\sigma\left(\mathbf{w}_1^{n-j}, \mathbf{z}\right)\right)} W_{n-j+1}\right] M_{\sigma'_{W_i}\left(\sigma\left(\mathbf{w}_1^{i-1}, \mathbf{z}\right)\right)}\tilde{W}_i\sigma\left(\mathbf{w}_1^{i-1}, \mathbf{z}\right).\end{aligned} \tag{60}$$

Furthermore, by (56) and (57), and the sub-multiplicativity of the operator norm, one obtains the first inequality

$$\begin{aligned}\left\|\partial_{W_i}\sigma\left(\mathbf{w}_1^n,\mathbf{z}\right)\right\| &\leq \sqrt{d_{i-1}}(\|\sigma\|_\infty+|\mathbf{z}|)\left\|\sigma'\right\|_\infty^{n-i+1}\prod_{j=1}^{n-i}\|W_{n-j+1}\|\\ &\leq \sqrt{D}(1+|x|)(1+|||\sigma|||)^{n-i+2}|\theta|^{n-i},\end{aligned}$$

since, by definition, $\sigma(\mathbf{w}_1^0,\mathbf{z})=\mathbf{z}$. In addition, due to the properties of the Euclidean norm,

$$\begin{aligned}|\partial_\theta f(\theta,\mathbf{z})|^2 =&|\partial_\phi f(\theta,\mathbf{z})|^2+\sum_{i=1}^n|\partial_{W_i}f(\theta,\mathbf{z})|^2\\ \leq& d_n\|\sigma\|_\infty^2+D|\phi|^2\sum_{i=1}^n(1+|x|)^2(1+|||\sigma|||)^{2(n-i+2)}|\theta|^{2(n-i)}\\ \leq& D(1+|x|)^2(1+|||\sigma|||)^{2(n+1)}\sum_{i=0}^n|\theta|^{2(n-i+1)}\\ \leq& D(n+1)(1+|x|)^2(1+|||\sigma|||)^{2(n+1)}(1+|\theta|^{2n}).\end{aligned}$$

Finally, the subadditivity of the square root function yields that

$$|\partial_\theta f(\theta,\mathbf{z})|\leq D^{1/2}\sqrt{n+1}(1+|x|)(1+|||\sigma|||)^{n+1}(1+|\theta|^n)$$

which completes the proof. □

Corollary A.7. *Let $\theta,\theta'\in\Theta$ and $x\in\mathbb{R}^m$ be such that $\theta=(\phi,\mathbf{w}_1^n)$, $\theta=(\phi',\mathbf{w'}_1^n)$ and $x=(\mathbf{z},y)$, where $\mathbf{w}_1^n,\mathbf{w'}_1^n\in\bigoplus_{i=1}^n Lin\left(\mathbb{R}^{d_{i-1}},\mathbb{R}^{d_i}\right)$, $\phi,\phi'\in\left(\mathbb{R}^{d_n}\right)^*$ and $x\in\mathbb{R}^m$ are arbitrary. Then, by Lemma A.6, for $t\in[0,1]$ and $i=1,\ldots,n$, follows that*

$$\begin{aligned}\left\|\partial_{\mathbf{w}_1^i}\sigma((1-t)\mathbf{w}_1^i+t\mathbf{w'}_1^i,\mathbf{z})\right\|^2 \leq&\sum_{j=1}^i\left\|\partial_{W_j}\sigma((1-t)\mathbf{w}_1^i+t\mathbf{w'}_1^i,\mathbf{z})\right\|^2\\ \leq& D(1+|x|)^2\sum_{j=1}^i(1+|||\sigma|||)^{2(n-j+2)}|(1-t)\theta+t\theta'|^{2(n-j)}\\ \leq& nD(1+|x|)^2(1+|||\sigma|||)^{2(n+1)}(1+|\theta|+|\theta'|)^{2(n-1)}\end{aligned}$$

which leads to the uniform estimate

$$\begin{aligned}\left|\sigma\left(\mathbf{w}_1^i,\mathbf{z}\right)-\sigma\left(\mathbf{w'}_1^i,\mathbf{z}\right)\right| \leq&\sup_{t\in[0,1]}\|\partial_{\mathbf{w}_1^i}\sigma((1-t)\mathbf{w}_1^i+t\mathbf{w'}_1^i,\mathbf{z})\||\mathbf{w}_1^i-\mathbf{w'}_1^i|\\ \leq& D^{1/2}\sqrt{n}(1+|x|)(1+|||\sigma|||)^{n+1}(1+|\theta|+|\theta'|)^{n-1}|\mathbf{w}_1^i-\mathbf{w'}_1^i|\\ \leq& D^{1/2}\sqrt{n}(1+|x|)(1+|||\sigma|||)^{n+1}(1+|\theta|+|\theta'|)^{n-1}|\theta-\theta'|\end{aligned}$$

$i=1,\ldots,n$.

Lemma A.8. *Let $\theta,\theta'\in\Theta$ and $x\in\mathbb{R}^m$ be such that $\theta=(\phi,\mathbf{w}_1^n)$, $\theta=(\phi',\mathbf{w'}_1^n)$ and $x=(\mathbf{z},y)$, where $\mathbf{w}_1^n,\mathbf{w'}_1^n\in\bigoplus_{i=1}^n Lin\left(\mathbb{R}^{d_{i-1}},\mathbb{R}^{d_i}\right)$, $\phi,\phi'\in\left(\mathbb{R}^{d_n}\right)^*$ and $x\in\mathbb{R}^m$ are arbitrary. Then, for $i=1,\ldots,n$, we have*

$$\left\|\partial_{W_i}\sigma(\mathbf{w}_1^n,\mathbf{z})-\partial_{W_i}\sigma(\mathbf{w'}_1^n,\mathbf{z})\right\|=2\sqrt{n}D(1+|x|)^2(1+|||\sigma|||)^{2n-i+4}(1+|\theta|+|\theta'|)^{2n-i}|\theta-\theta'|.$$

Proof. Let $i\in\{1,\ldots,n\}$ be arbitrary and fixed. By the definition of $\sigma(\mathbf{w}_1^n,\mathbf{z})$, for $k<n$, $\sigma\left(\mathbf{w}_1^{k+1},\mathbf{z}\right)=\sigma_{W_{k+1}}\circ\sigma\left(\mathbf{w}_1^k,\mathbf{z}\right)$. Hence, for $i\leq k<n$,

$$\partial_{W_i}\sigma\left(\mathbf{w}_1^{k+1},\mathbf{z}\right)=M_{\sigma'_{W_{k+1}}\left(\sigma\left(\mathbf{w}_1^k,\mathbf{z}\right)\right)}W_{k+1}\partial_{W_i}\sigma\left(\mathbf{w}_1^k,\mathbf{z}\right)$$

which implies that

$$\begin{aligned}\left\|\partial_{W_i}\sigma\left(\mathbf{w}_1^{k+1},\mathbf{z}\right)-\partial_{W_i}\sigma\left(\mathbf{w'}_1^{k+1},\mathbf{z}\right)\right\| \leq&\left\|M_{\sigma'_{W_{k+1}}\left(\sigma\left(\mathbf{w}_1^k,\mathbf{z}\right)\right)}-M_{\sigma'_{W'_{k+1}}\left(\sigma\left(\mathbf{w'}_1^k,\mathbf{z}\right)\right)}\right\|\\ &\times\|W_{k+1}\|\left\|\partial_{W_i}\sigma\left(\mathbf{w}_1^k,\mathbf{z}\right)\right\|\\ &+\left\|M_{\sigma'_{W'_{k+1}}\left(\sigma\left(\mathbf{w'}_1^k,\mathbf{z}\right)\right)}\right\|\|W_{k+1}-W'_{k+1}\|\\ &\times\left\|\partial_{W_i}\sigma\left(\mathbf{w}_1^k,\mathbf{z}\right)\right\|\end{aligned}$$

$$
\begin{aligned}
&+\left\|M_{\sigma'_{W'_{k+1}}\left(\sigma\left(\mathbf{w'}_1^k,\mathbf{z}\right)\right)}\right\| \\
&\times\|W'_{k+1}\|\left\|\partial_{W_i}\sigma\left(\mathbf{w}_1^k,\mathbf{z}\right)-\partial_{W_i}\sigma\left(\mathbf{w'}_1^k,\mathbf{z}\right)\right\| \\
\leq&\|\sigma'\|_\infty|\mathbf{w'}_1^n|\left\|\partial_{W_i}\sigma\left(\mathbf{w}_1^k,\mathbf{z}\right)-\partial_{W_i}\sigma\left(\mathbf{w'}_1^k,\mathbf{z}\right)\right\| \\
&+\left\|\partial_{W_i}\sigma\left(\mathbf{w}_1^k,\mathbf{z}\right)\right\| \\
&\times\left[\left\|M_{\sigma'_{W_{k+1}}\left(\sigma\left(\mathbf{w}_1^k,\mathbf{z}\right)\right)}-M_{\sigma'_{W'_{k+1}}\left(\sigma\left(\mathbf{w'}_1^k,\mathbf{z}\right)\right)}\right\||\mathbf{w}_1^n|\right. \\
&\qquad\left.+\|\sigma'\|_\infty|\mathbf{w}_1^n-\mathbf{w'}_1^n|\right]
\end{aligned}
$$

holds for the corresponding operator norms. Further, for $i=1,\ldots,n$ and by taking into consideration Corollary A.7, one obtains that

$$
\begin{aligned}
\left\|M_{\sigma'_{W_i}\left(\sigma\left(\mathbf{w}_1^{i-1},\mathbf{z}\right)\right)}-M_{\sigma'_{W'_i}\left(\sigma\left(\mathbf{w'}_1^{i-1},\mathbf{z}\right)\right)}\right\|&=\left\|M_{\sigma'_{W_i}\left(\sigma\left(\mathbf{w}_1^{i-1},\mathbf{z}\right)\right)-\sigma'_{W'_i}\left(\sigma\left(\mathbf{w'}_1^{i-1},\mathbf{z}\right)\right)}\right\| \\
&\leq\left\|\sigma'_{W_i}\left(\sigma\left(\mathbf{w}_1^{i-1},\mathbf{z}\right)\right)-\sigma'_{W'_i}\left(\sigma\left(\mathbf{w'}_1^{i-1},\mathbf{z}\right)\right)\right\|_\infty \\
&\leq\|\sigma'\|_{\text{Lip}}\left(\|W_i\|\left|\sigma\left(\mathbf{w}_1^{i-1},\mathbf{z}\right)-\sigma\left(\mathbf{w'}_1^{i-1},\mathbf{z}\right)\right|\right. \\
&\qquad\left.+\|W_i-W'_i\|\left|\sigma\left(\mathbf{w'}_1^{i-1},\mathbf{z}\right)\right|\right) \\
&\leq D^{1/2}\sqrt{n}(1+|||\sigma|||)^{n+2}(1+|x|)(1+|\theta|+|\theta'|)^n|\theta-\theta'|
\end{aligned}
\tag{61}
$$

which is uniform in i. Combining these with inequality (59) in Lemma A.6, for $i\leq k<n$, one obtains the following recursive estimate

$$
\left\|\partial_{W_i}\sigma\left(\mathbf{w}_1^{k+1},\mathbf{z}\right)-\partial_{W_i}\sigma\left(\mathbf{w'}_1^{k+1},\mathbf{z}\right)\right\|\leq A\left\|\partial_{W_i}\sigma\left(\mathbf{w}_1^k,\mathbf{z}\right)-\partial_{W_i}\sigma\left(\mathbf{w'}_1^k,\mathbf{z}\right)\right\|+BA^{n+k-i+1}, \tag{62}
$$

where

$$
\begin{aligned}
A&=(1+|||\sigma|||)(1+|\theta|+|\theta'|) \\
B&=2\sqrt{n}D(1+|x|)^2(1+|||\sigma|||)^4|\theta-\theta'|.
\end{aligned}
$$

By induction, for $i=1,\ldots,n$, one deduces that

$$
\left\|\partial_{W_i}\sigma\left(\mathbf{w}_1^n,\mathbf{z}\right)-\partial_{W_i}\sigma\left(\mathbf{w'}_1^n,\mathbf{z}\right)\right\|\leq A^{n-i}\left\|\partial_{W_i}\sigma\left(\mathbf{w}_1^i,\mathbf{z}\right)-\partial_{W_i}\sigma\left(\mathbf{w'}_1^i,\mathbf{z}\right)\right\|+(n-i)BA^{2n-i}. \tag{63}
$$

Using basic properties of the operator norm and inequality (55), for $i=n$, yields that

$$
\begin{aligned}
\left|\partial_{W_i}\sigma\left(\mathbf{w}_1^i,\mathbf{z}\right)(\tilde{W}_i)-\partial_{W_i}\sigma\left(\mathbf{w'}_1^i,\mathbf{z}\right)(\tilde{W}_i)\right|=&\left|M_{\sigma'_{W_i}\left(\sigma\left(\mathbf{w}_1^{i-1},\mathbf{z}\right)\right)}\tilde{W}_i\sigma\left(\mathbf{w}_1^{i-1},\mathbf{z}\right)\right. \\
&\left.-M_{\sigma'_{W'_i}\left(\sigma\left(\mathbf{w'}_1^{i-1},\mathbf{z}\right)\right)}\tilde{W}_i\sigma\left(\mathbf{w'}_1^{i-1},\mathbf{z}\right)\right| \\
\leq&\left\|M_{\sigma'_{W_i}\left(\sigma\left(\mathbf{w}_1^{i-1},\mathbf{z}\right)\right)}\right\|\left|\sigma\left(\mathbf{w}_1^{i-1},\mathbf{z}\right)-\sigma\left(\mathbf{w'}_1^{i-1},\mathbf{z}\right)\right||\tilde{W}_i| \\
&+\left\|M_{\sigma'_{W_i}\left(\sigma\left(\mathbf{w}_1^{i-1},\mathbf{z}\right)\right)}-M_{\sigma'_{W'_i}\left(\sigma\left(\mathbf{w'}_1^{i-1},\mathbf{z}\right)\right)}\right\| \\
&\times\left|\sigma\left(\mathbf{w'}_1^{i-1},\mathbf{z}\right)\right||\tilde{W}_i|
\end{aligned}
$$

which, due to Corollary A.7 and (61), implies that

$$
\begin{aligned}
\left\|\partial_{W_i}\sigma\left(\mathbf{w}_1^i,\mathbf{z}\right)-\partial_{W_i}\sigma\left(\mathbf{w'}_1^i,\mathbf{z}\right)\right\|\leq&\left\|M_{\sigma'_{W_i}\left(\sigma\left(\mathbf{w}_1^{i-1},\mathbf{z}\right)\right)}\right\|\left|\sigma\left(\mathbf{w}_1^{i-1},\mathbf{z}\right)-\sigma\left(\mathbf{w'}_1^{i-1},\mathbf{z}\right)\right| \\
&+\left\|M_{\sigma'_{W_i}\left(\sigma\left(\mathbf{w}_1^{i-1},\mathbf{z}\right)\right)}-M_{\sigma'_{W'_i}\left(\sigma\left(\mathbf{w'}_1^{i-1},\mathbf{z}\right)\right)}\right\|\left|\sigma\left(\mathbf{w'}_1^{i-1},\mathbf{z}\right)\right| \\
\leq&BA^n.
\end{aligned}
$$

Finally, combine this estimate with (63) yields that

$$
\left\|\partial_{W_i}\sigma(\mathbf{w}_1^n,\mathbf{z})-\partial_{W_i}\sigma(\mathbf{w'}_1^n,\mathbf{z})\right\|\leq(n-i+1)BA^{2n-i}
$$

$$= 2\sqrt{n}D(1+|x|)^2(1+|||\sigma|||)^{2n-i+4}(1+|\theta|+|\theta'|)^{2n-i}|\theta-\theta'|$$

which completes the proof. □

Lemma A.9. *Let $x=(\mathbf{z},y)$, where $\mathbf{z}\in\mathbb{R}^{m-1}$ and $y\in\mathbb{R}$ are arbitrary. Then, for any $\theta,\theta'\in\Theta$,*

$$|\partial_\theta f(\theta,\mathbf{z})-\partial_\theta f(\theta',\mathbf{z})|\leq 4(n+1)D(1+|x|)^2(1+|||\sigma|||)^{2n+3}(1+|\theta|+|\theta'|)^{2n}|\theta-\theta'|.$$

Proof. For the Euclidean norm of the partial derivative of the regression function with respect to the learning parameter, we have

$$|\partial_\theta f(\theta,\mathbf{z})|^2=|\sigma(\mathbf{w}_1^n,\mathbf{z})|^2+\sum_{i=1}^n|\phi\circ\partial_{W_i}\sigma(\mathbf{w}_1^n,\mathbf{z})|^2$$

and thus we have

$$|\partial_\theta f(\theta,\mathbf{z})-\partial_\theta f(\theta',\mathbf{z})|^2=|\sigma(\mathbf{w}_1^n,\mathbf{z})-\sigma(\mathbf{w'}_1^n,\mathbf{z})|^2+\sum_{i=1}^n|\phi\circ\partial_{W_i}\sigma(\mathbf{w}_1^n,\mathbf{z})-\phi'\circ\partial_{W_i}\sigma(\mathbf{w'}_1^n,\mathbf{z})|^2.$$

Using Lemma A.6 and A.8, one deduces that

$$\begin{aligned}\left|\phi\circ\partial_{W_i}\sigma(\mathbf{w}_1^n,\mathbf{z})-\phi'\circ\partial_{W_i}\sigma(\mathbf{w'}_1^n,\mathbf{z})\right|^2\leq&2\Big(|\phi|^2\left\|\partial_{W_i}\sigma(\mathbf{w}_1^n,\mathbf{z})-\partial_{W_i}\sigma(\mathbf{w'}_1^n,\mathbf{z})\right\|^2\\&+|\phi-\phi'|^2\left\|\partial_{W_i}\sigma(\mathbf{w'}_1^n,\mathbf{z})\right\|^2\Big)\\\leq&8nD^2(1+|||\sigma|||)^{2(2n-i+4)}(1+|\theta|+|\theta'|)^{2+4n-2i}|\theta-\theta'|^2\\&+2D(1+|x|)^2(1+|||\sigma|||)^{2(n-i+2)}(1+|\theta|+|\theta'|)^{2n-2i}|\\&\times\theta-\theta'|^2\\\leq&16nD^2(1+|x|)^4(1+|||\sigma|||)^{2(2n-i+4)}(1+|\theta|+|\theta'|)^{2+4n-2i}\\&\times|\theta-\theta'|^2.\end{aligned}$$

moreover by, Corollary A.7, for the first term, we have

$$|\sigma(\mathbf{w}_1^n,\mathbf{z})-\sigma(\mathbf{w'}_1^n,\mathbf{z})|^2\leq Dn(1+|x|)^2(1+|||\sigma|||)^{2(n+1)}(1+|\theta|+|\theta'|)^{2(n-1)}|\theta-\theta'|^2.$$

Hence

$$|\partial_\theta f(\theta,\mathbf{z})-\partial_\theta f(\theta',\mathbf{z})|^2\leq 16(n+1)^2D^2(1+|x|)^4(1+|||\sigma|||)^{2(2n+3)}(1+|\theta|+|\theta'|)^{4n}|\theta-\theta'|^2.$$

□

The next Proposition asserts that the growth condition 2 holds with

$$K(x)=4D\sqrt{n+1}(1+|x|)^2(1+|||\sigma|||)^{n+2}$$

whenever $r\geq\frac{n+3}{2}$.

Proposition A.10. *For any $\theta\in\Theta$ and $x\in\mathbb{R}^m$,*

$$|G(\theta,x)|\leq 4D\sqrt{n+1}(1+|x|)^2(1+|||\sigma|||)^{n+2}(1+|\theta|^{n+1}).$$

Proof. By Lemma A.6, for arbitrary $x\in\mathbb{R}^m$ and $\theta\in\mathbb{R}^d$, one calculates

$$\begin{aligned}|G(\theta,x)|=\|G(\theta,x)\|&=2|y-f(\theta,\mathbf{z})||\partial_\theta f(\theta,\mathbf{z})|\\&\leq 2(|y|+|f(\theta,\mathbf{z})|)|\partial_\theta f(\theta,\mathbf{z})|\\&\leq 2(1+|x|)D^{1/2}(1+|||\sigma|||)(1+|\theta|)|\partial_\theta f(\theta,\mathbf{z})|\\&\leq 4D\sqrt{n+1}(1+|x|)^2(1+|||\sigma|||)^{n+2}(1+|\theta|^{n+1})\end{aligned}$$

since $|\theta|+|\theta|^n\leq 1+|\theta|^{n+1}$, for any $n\geq 1$. □

The next Proposition states that Assumption 1 is satisfied with $\rho=3$, $q-1=\max(2n+1,2r)$ and

$$L_1=16(1+\eta)(2r+1)(n+1)D^{3/2}(1+|||\sigma|||)^{2n+4}.$$

Proposition A.11 (*Link to Assumption 1 and Proposition 2.7*)**.** *For any $\theta\in\Theta$ and $x\in\mathbb{R}^m$,*

$|H(\theta,x)-H(\theta',x)|\leq 16(1+\eta)(2r+1)(n+1)D^{3/2}(1+|x|)^3(1+|||\sigma|||)^{2n+4}(1+|\theta|+|\theta'|)^{q-1}|\theta-\theta'|$
where $q-1=\max(2n+1,2r)$.

Proof of Proposition 4.1. In view of Lemmas A.6 and A.9 and Corollary A.7, one obtains for *the first term that satisfies Assumption 1* since

$$\begin{aligned}\frac{1}{2}|G(\theta,x)-G(\theta',x)| \leq & |y-f(\theta,\mathbf{z})||\partial_\theta f(\theta,\mathbf{z})-\partial_\theta f(\theta',\mathbf{z})|+|f(\theta,\mathbf{z})-f(\theta',\mathbf{z})||\partial_\theta f(\theta',\mathbf{z})| \\ \leq & 4(n+1)D^{3/2}(1+|x|)^3(1+|||\sigma|||)^{2n+4}(1+|\theta|+|\theta'|)^{2n+1}|\theta-\theta'| \\ & +2(n+1)(1+|x|)^2(1+|||\sigma|||)^{2n+2}(1+|\theta|+|\theta'|)^{2n}|\theta-\theta'| \\ \leq & 8(n+1)D^{3/2}(1+|x|)^3(1+|||\sigma|||)^{2n+4}(1+|\theta|+|\theta'|)^{2n+1}|\theta-\theta'|,\end{aligned}$$

which completes the proof. □