Polygonal Unadjusted Langevin Algorithms: Creating stable and efficient adaptive algorithms for neural networks

Dong-Young Lim DLIM@UNIST.AC.KR
Department of Industrial Engineering
Artificial Intelligence Graduate School
UNIST
Ulsan, 44919, South Korea

Sotirios Sabanis S.SABANIS@ED.AC.UK
School of Mathematics
The University of Edinburgh
Edinburgh, EH9 3FD, UK
The Alan Turing Institute
London NW1 2DB, UK

Abstract

We present a new class of Langevin based algorithms, which overcomes many of the known shortcomings of popular adaptive optimizers that are currently used for the fine tuning of deep learning models. Its underpinning theory relies on recent advances of Euler's polygonal approximations for stochastic differential equations (SDEs) with monotone coefficients. As a result, it inherits the stability properties of tamed algorithms, while it addresses other known issues, e.g. vanishing gradients in deep learning. In particular, we provide a nonasymptotic analysis and full theoretical guarantees for the convergence properties of an algorithm of this novel class, which we named THεO POULA (or, simply, TheoPouLa). Finally, several experiments are presented with different types of deep learning models, which show the superior performance of TheoPouLa over many popular adaptive optimization algorithms.

Keywords: Stochastic optimization, nonconvex optimization, non-asymptotic estimates, taming technique, Euler's polygonal approximation

1. Introduction

Modern machine learning models including deep neural networks are successfully trained when they are finely tuned via the optimization of their associated loss functions. Two aspects of such optimization tasks pose significant challenges, namely the nonconvex nature of loss functions and the highly nonlinear features of many types of neural networks. Moreover, the analysis in Lovas et al. (2020) shows that the gradients of such nonconvex loss functions typically grow faster than linearly and are only locally Lipschitz continuous. Naturally, stability issues are observed, which are known as the 'exploding gradient' phenomenon (Bengio et al. 1994; Pascanu et al. 2013), when vanilla stochastic gradient descent (SGDs) or certain types of adaptive algorithms are used for fine tuning. In addition, the sparsity of gradients of neural networks is another challenging issue, which is extensively studied in the literature. For example, momentum methods and adaptive learning rate methods such as AdaGrad (Duchi et al. 2011), RMSProp (Tieleman and Hinton 2012) and Adam (Kingma

and Ba 2015) have been developed to tackle this problem by diagonally scaling the gradient by some function taking the past gradients, also known as preconditioner.

Langevin based algorithms have been another important stream of literature on stochastic optimization. They are built on the theoretical fact that the Langevin stochastic differential equation (7), under mild conditions, converges to a unique invariant measure with a very attractive property. The measure concentrates around the global minimizers of the objective function as $\beta \to \infty$, see Hwang (1980), even in the case of nonconvex potentials. As a result, the global convergence of the stochastic gradient Langevin dynamics (SGLD) and its variants has been extensively studied in a nonconvex setting, see, e.g., Raginsky et al. (2017); Xu et al. (2018); Erdogdu et al. (2018); Brosse et al. (2018); Lovas et al. (2020). Moreover, it is worth noting that Langevin based algorithms have been a key element in Bayesian statistics and in Markov Chain Monte Carlo (MCMC) theory, see, e.g., Roberts and Tweedie (1996); Durmus and Moulines (2017); Dalalyan (2017); Brosse et al. (2019); Welling and Teh (2011); Deng et al. (2020a,b).

Motivated by the aforementioned developments in the field, we propose a new class of stochastic gradient Langevin algorithms that addresses several challenges in deep learning. Its underpinning theory relies on recent advances of Euler's polygonal approximations for stochastic differential equations (SDEs) with monotone coefficients, which originate from the articles Krylov (1985) and Krylov (1990). We name this new class as *polygonal unadjusted Langevin algorithms*. Mathematically, it is described as follows: Given an i.i.d. sequence of random variables $\{X_n\}_{n\geq 0}$ of interest, which typically represent available data, the algorithm follows

$$\theta^{\lambda}_{n+1} := \theta^{\lambda}_n - \lambda H_\lambda(\theta^{\lambda}_n, X_{n+1}) + \sqrt{2\lambda\beta^{-1}}\xi_{n+1}, \tag{1}$$

where $n \in \mathbb{N}$, $\theta^{\lambda}_0 := \theta_0$, θ_0 is an $\mathbb{R}^d$-valued random variable, $\lambda > 0$ denotes the step size of the algorithm, $\beta > 0$ is the so-called inverse temperature, $(\xi_n)_{n\in\mathbb{N}}$ is an $\mathbb{R}^d$-valued Gaussian process with i.i.d. components and $H_\lambda : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ satisfies the following three properties:

1. For every $\lambda > 0$, there exist constants $K_\lambda > 0$ and $\rho_1 \geq 0$ such that

$$|H_\lambda(\theta, x)| \leq K_\lambda(1 + |x|)^{\rho_1}(1 + |\theta|) \text{ for every } \theta \in \mathbb{R}^d \text{ and } x \in \mathbb{R}^m.$$

2. There exist constants $\gamma \geq 1/2$, $K_2 > 0$ and ρ_2, $\rho_3 \geq 0$ such that for all $\lambda > 0$,

$$|H_\lambda(\theta, x) - H(\theta, x)| \leq \lambda^\gamma K_2(1 + |x|)^{\rho_2}(1 + |\theta|)^{\rho_3} \text{ for every } \theta \in \mathbb{R}^d \text{ and } x \in \mathbb{R}^m,$$

   where H is the (unbiased) stochastic gradient of the objective function of the optimization problem.

3. There exist constants λ_{max} and $\delta \in \{1, 2\}$ such that for any $\lambda \leq \lambda_{max}$,

$$\liminf_{|\theta|\to\infty} \mathbb{E}\left[\langle \frac{\theta}{|\theta|^\delta}, H_\lambda(\theta, X_0)\rangle - \frac{2\lambda}{|\theta|^\delta}|H_\lambda(\theta, X_0)|^2\right] > 0.$$

One obtains our new algorithm TheoPouLa by considering the case where $H_\lambda(\theta, x)$ is the vector with entries $H^{(i)}_{\lambda,c}(\theta, x)$ as given by (9), for $i \in \{1, \ldots, d\}$. The flexibility of Euler's

polygonal approximations allows TheoPouLa to combine an element-wise taming function and a boosting function, which effectively resolve the 'exploding gradient' and 'vanishing gradient' problems that are frequently observed in deep learning. More specifically, the element-wise taming function is proposed to control the super-linearly growing stochastic gradient[1] in high-dimensional optimization problems, which successfully extends the taming technique of TUSLA (Lovas et al. 2020). In the literature, the taming techniques have been widely studied in the construction of stable numerical approximations. For example, see Hutzenthaler et al. (2012); Sabanis (2013, 2016) for nonlinear SDEs and Brosse et al. (2019); Sabanis and Zhang (2019) for MCMC algorithms. Furthermore, the boosting function is introduced to address the sparsity in deep learning by adaptively adjusting the stepsize of the algorithm in a region where the loss function is flat, i.e., the gradient is small. As a result, the flat gradients can increase by up to some point that is controlled by the hyperparameter of TheoPouLa, denoted by ε. Moreover, properly scaled isotropic Gaussian noise is added at each iteration since TheoPouLa is essentially a type of Langevin based algorithms. Hence, its name is formed from the above description, so called Tamed Hybrid ε-Order POlygonal Unadjusted Langevin Algorithm (THεO POULA or TheoPouLa). We note that TheoPouLa and TUSLA (Lovas et al. 2020) satisfy the above three properties with $\delta = 2$ and $\gamma = 1/2$, whereas TULA (Brosse et al. 2019) satisfies them with $\delta = \gamma = 1$ as it assumes only deterministic gradients (and thus the i.i.d. data sequence reduces to a constant).

In Section 2, the precise formula of TheoPouLa and its full detailed analysis (including its convergence properties) are given. Furthermore, we provide in Section 3.2 extensive numerical experiments which demonstrate remarkable empirical performance of TheoPouLa on real-world datasets such as CIFAR10 and CIFAR100 for image classification, and the Penn Treebank for language modeling. Section 3.3 investigates the effect of the key hyperparameters of TheoPouLa on its performance and Section 3.4 presents additional experiments to support the effectiveness of the boosting function. All the proofs of main results in Section 2 are provided in Section 4.

1.1 Related work: Langevin based algorithms and adaptive learning rate methods

In this paper, we focus on reviewing the literature studying Langevin based algorithms for optimization problems. We refer to Welling and Teh (2011); Ahn et al. (2021); Chen et al. (2014); Deng et al. (2020a,b); Zhang et al. (2020) and references therein for recent progress on MCMC algorithms and Bayesian neural networks. Most research on Langevin based algorithms for nonconvex optimization in the literature has been focused on theoretical aspects. Raginsky et al. (2017) demonstrated the links between Langevin based algorithms and stochastic optimization in neural networks, stimulating further the development and analysis of such algorithms. Xu et al. (2018) analyzed the global convergence of gradient Langevin dynamics (GLD), stochastic gradient Langevin dynamics (SGLD) and stochastic variance reduced gradient Langevin dynamics (SVRG-LD). The incorporation of dependent data streams in the analysis of SGLD algorithms has been achieved in Barkhagen

1. Hutzenthaler et al. (2012) show that the Euler discretization with super-linearly grwoing coefficients could diverge to infinity in finite time.

et al. (2021) and in Chau et al. (2019), and local conditions have been studied in Zhang et al. (2019). Recently, TUSLA in Lovas et al. (2020) has been proposed based on a new generation of tamed Euler approximations for stochastic differential equations (SDEs) with monotone coefficients in nonconvex optimization problems. Despite their elegant theoretical results, the use of Langevin based algorithms for training deep learning models has been limited in practice as their empirical performance lacked behind in comparison to popular adaptive learning rate methods.

Adaptive learning rate methods such as AdaGrad (Duchi et al. 2011), RMSProp (Tieleman and Hinton 2012), and Adam (Kingma and Ba 2015) have been successfully applied to neural network models due to their fast training speed. In particular, Adam-type optimizers can be generally written as follows, for $n \in \mathbb{N}_0$,

$$\theta_{n+1} \quad = \quad \theta_n - \lambda \frac{m_n}{\varepsilon + \sqrt{V}_n} \tag{2}$$

where $m_n = \phi_n(H_1, \cdots, H_n)$, $V_n = \psi_n(H_1, \cdots, H_n)$ is a preconditioning matrix, $H_i := H(\theta_i, X_i)$ is the stochastic gradient evaluated at the i-th iteration, λ is the learning rate and all operations are applied element-wise. Table 1 provides the details for some of popular stochastic optimization methods with corresponding averaging functions ϕ_n and ψ_n. Since

Table 1: Summary of stochastic optimization methods within the general framework. Note that $\widehat{v}_n = \max\{\widehat{v}_{n-1}, v_n\}$ is defined as $v_n = (1-\beta_2)v_{n-1} + \beta_2 H_n^2$.

	SGD	RMSProp	Adam	AMSGrad
$\phi_n :=$	H_n	H_n	$(1-\beta_1)\sum_{i=1}^n \beta_1^{n-i} H_i$	$(1-\beta_1)\sum_{i=1}^n \beta_1^{n-i} H_i$
$\psi_n :=$	$\mathbb{I}_n$	$(1-\beta_2)diag(\sum_{i=1}^n \beta_2^{n-i} H_i^2)$	$(1-\beta_2)diag(\sum_{i=1}^n \beta_2^{n-i} H_i^2)$	$diag(\widehat{v}_n)$

the appearance of Adam, a large number of variants of Adam-type optimizers have been proposed to address the theoretical and practical challenges of Adam by suggesting a new preconditioner, V_n in (2), to scale the stochastic gradient. For example, Reddi et al. (2018) provided a simple example that demonstrates the non-convergence issue of Adam and proposed a simple modification, called AMSGrad, to solve the problem. Chen et al. (2019) discussed the convergence of Adam-type optimizers in a nonconvex setting. RAdam to rectify the variance of adaptive learning rate has been proposed in Liu et al. (2020). Wilson et al. (2017) revealed that the generalization ability of adaptive learning rate methods is worse than a global learning method like SGD. AdaBound of Luo et al. (2019) attempts to overcome the drawback by employing dynamic bounds on learning rates. Recently, AdaBelief (Zhuang et al. 2020) and AdamP (Heo et al. 2021) demonstrated their fast convergence and good generalization via extensive experiments. Nevertheless, these (and other) adaptive learning rate methods have an obvious theoretical drawback as they are only guaranteed to converge to a stationary point, which can be a local minimum or even a saddle point in nonconvex settings. In addition, the theoretical results require strong assumptions such as the global Lipschitz continuity and boundedness conditions on the stochastic gradient. One should note that none of these two assumptions hold true in a typical optimization

problem involving neural networks. This is particularly evident in complex neural network architectures.

1.2 Our contributions

The newly proposed algorithm, TheoPouLa, combines both advantages: global convergence in Langevin based algorithms and powerful empirical performance in adaptive learning rate methods. To the best of the authors' knowledge, our algorithm is the first Langevin based algorithm to achieve a comparable (or even better) empirical performance in deep learning tasks compared to popular stochastic optimization methods such as SGD, Adam, AMSGrad, RMSProp, AdaBound and AdaBelief. The major strengths of our work over related algorithms are summarized as follows:

- We provide a global convergence analysis of TheoPouLa in Wasserstein 1 and 2 distances when the stochastic gradient of the objective function is locally Lipscthiz continuous. Moreover, a non-asymptotic estimate for the expected excess risk of the algorithm is derived.

- Polygonal unadjusted Langevin algorithms significantly extend possible approximations for the drift term of the Langevin SDE, which allows the algorithm to deal with the exploding and vanishing gradient problems. In particular, TheoPouLa achieves a stable and fast training process due to the element-wise taming technique and boosting function, which are theoretically well-designed for the algorithm to adaptively take a desirable stepsize. Furthermore, the effectiveness of both taming and boosting functions is confirmed through several empirical experiments.

- While TheoPouLa behaves like adaptive learning rate methods in the early training phase, it takes an almost global learning rate near an optimal point. In other words, TheoPouLa is quickly switched from adaptive methods to SGD. As a result, it inherits the good generalization ability of SGD. Our experiments support this fact by showing that TheoPouLa outperforms the other optimization methods in *generalization* measured by test accuracy for various deep learning tasks.

2. New Algorithm: THεO POULA

We propose a new stochastic optimization algorithm by combining ideas from taming methods specifically designed to approximate Langevin SDEs with a hybrid approach based on recent advances of polygonal Euler approximations. The latter is achieved by identifying a suitable boosting function (of order $\varepsilon \ll 1$) to efficiently deal with the sparsity of the stochastic gradients of neural networks. The novelty of our algorithm is to utilize a taming function and a boosting function instead of designing a new preconditioner V_n in (2) as in Adam-type optimizers.

We proceed with the necessary preliminary information, main assumptions and formal introduction of the new algorithm.

2.1 Preliminaries and Assumptions

Let $(\Omega, \mathcal{F}, P)$ be a probability space. We denote by $\mathbb{E}[X]$ the expectation of a random variable X. Fix an integer $k \geq 1$. For an $\mathbb{R}^k$-valued random variable X, its law on $\mathcal{B}(\mathbb{R}^k)$, i.e. the Borel sigma-algebra of $\mathbb{R}^k$, 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 $\mu \in \mathcal{P}(\mathbb{R}^k)$ and for a non-negative measurable $f : \mathbb{R}^k \to \mathbb{R}$, the notation $\mu(f) := \int_{\mathbb{R}^k} f(\theta)\mu(\mathrm{d}\theta)$ 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 $\mu, \nu \in \mathcal{P}(\mathbb{R}^k)$, let $\mathcal{C}(\mu, \nu)$ denote the set of probability measures ζ on $\mathcal{B}(\mathbb{R}^{2k})$ 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

$$W_p(\mu, \nu) := \inf_{\zeta \in \mathcal{C}(\mu,\nu)} \left(\int_{\mathbb{R}^k} \int_{\mathbb{R}^k} |\theta - \theta'|^p \zeta(\mathrm{d}\theta \mathrm{d}\theta') \right)^{1/p} \tag{3}$$

for $\mu, \nu \in \mathcal{P}(\mathbb{R}^k)$. Let $(X_n)_{n\in\mathbb{N}_0}$ be a sequence of i.i.d. $\mathbb{R}^m$-valued random variables generating the filtration $(\mathcal{G}_n)_{n\in\mathbb{N}_0}$ and $(\xi_n)_{n\in\mathbb{N}_0}$ be an $\mathbb{R}^d$-valued Gaussian process with independent components. It is assumed throughout the paper that the random variable θ_0, $\mathcal{G}_\infty := \sigma\left(\cup_{n\in\mathbb{N}_0}\mathcal{G}_n\right)$, and $(\xi_n)_{n\in\mathbb{N}_0}$ are independent.

Let $F : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ be a continuously differentiable function such that $\mathbb{E}[|F(\theta, X_0)|] < \infty$, for all $\theta \in \mathbb{R}^d$, where X_0 is a given $\mathbb{R}^m$-valued random variable with probability law $\mathcal{L}(X_0)$. We then consider the following optimization problem

$$\min_{\theta\in\mathbb{R}^d} u(\theta) = \min_{\theta\in\mathbb{R}^d} \left(\mathbb{E}[F(\theta, X_0)] + \frac{\eta}{2(r+1)}|\theta|^{2(r+1)} \right) \tag{4}$$

where $r > 0$ and $\eta \in (0, 1)$. Assume that $u : \mathbb{R}^d \to \mathbb{R}$ is a continuously differentiable function and denote by $h := \nabla u$ its gradient.

In the context of fine tuning of neural networks, F represents the loss function for the task at hand and θ denotes the vector of the model's parameters. In particular, r is determined by the property of the model. More precisely, the regularization term $\frac{\eta}{2(r+1)}|\theta|^{2(r+1)}$ is added in order to guarantee a dissipativity condition (6), which is necessary for the convergence of optimization algorithms.

Remark 1. *For the reader who prefers to consider the optimization problem without the regularization term, i.e., $\eta = 0$, the dissipative condition* (6) *has to be additionally assumed as in the literature (Raginsky et al., 2017; Xu et al., 2018; Erdogdu et al., 2018). Then, the same analysis can be applied to obtain our main results without any additional effort. However, it is yet to be proven theoretically that such an assumption holds in general for neural networks and thus it becomes a case-by-case investigation. In other words, we present here the formal theoretical statement with the appropriate regularization term which covers all of these cases.*

We denote by $H : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ the stochastic gradient of the objective function, which is given by

$$H(\theta, x) := G(\theta, x) + \eta\theta|\theta|^{2r}, \tag{5}$$

where $G(\theta, x) := \nabla_\theta F(\theta, x)$ for all $x \in \mathbb{R}^m$, $\theta \in \mathbb{R}^d$. Note that $\eta = 0$ if dissipativity holds for G. In addition, it is assumed that $H(\theta, x)$ is an unbiased estimator of $h(\theta)$ for all $\theta \in \mathbb{R}^d, x \in \mathbb{R}^m$.

To derive our main results, we introduce the following assumptions. Let $q \in [1, \infty)$, $r \in [q/2, \infty)$, $\rho \in [1, \infty)$ be fixed. We then impose conditions on the initial value θ_0 and data process $(X_n)_{n\in\mathbb{N}_0}$. As it is common to use a weight initialization technique using the uniform or normal distribution, Assumption 1 is mild.

Assumption 1. *The process $(X_n)_{n\in\mathbb{N}_0}$ has a finite $16\rho(2r+1)$-th moment, i.e., $\mathbb{E}[|X_0|^{16\rho(2r+1)}] < \infty$ and the initial condition has a finite $16(2r+1)$-th moment, i.e., $\mathbb{E}[|\theta_0|^{16(2r+1)}] < \infty$.*

The second requirement is that G is locally Lipschitz continuous satisfying a polynomial growth condition, which is substantially weaker than a (globally) Lipscthiz continuity or a bounded condition in the existing literature.

Assumption 2. *There exists a constant $L_G > 0$ such that, for all $x \in \mathbb{R}^m$, $\theta, \theta' \in \mathbb{R}^d$,*

$$|G(\theta, x) - G(\theta', x)| \le L_G(1 + |x|)^\rho (1 + |\theta| + |\theta'|)^{q-1} |\theta - \theta'|.$$

Remark 2. *From Assumption 2, one obtains, for all $\theta \in \mathbb{R}^d$, $x \in \mathbb{R}^m$,*

$$|G(\theta, x)| \le K_G(x)(1 + |\theta|)^q,$$

where $K_G(x) = L_G(1 + |x|)^\rho + |G(0, x)|$.

Under Assumption 1 and 2, one can derive a dissipative condition for h, which is presented in the next remark.

Remark 3. *From Assumption 1 and 2, it can be shown that h satisfies the following dissipative condition, for all $\theta \in \mathbb{R}^d$,*

$$\langle \theta, h(\theta) \rangle \ge A|\theta|^2 - B, \tag{6}$$

where $A = 2^q \mathbb{E}[1 + K_G(X_0)]$, $B = 3(2^{q+1}\mathbb{E}[1 + K_G(X_0)])^{q+2}/\eta^{q+1}$.

Furthermore, under Assumption 1 and 2, the following proposition states that one can obtain an one-sided Lipschitz continuity condition for h. The proof of Proposition 4 can be found in (Lovas et al., 2020, Proposition 1).

Proposition 4. *Let Assumption 1 and 2 hold. Then, one obtains, for all θ, $\theta' \in \mathbb{R}^d$,*

$$\langle \theta - \theta', h(\theta) - h(\theta') \rangle \ge -L_R|\theta - \theta'|^2,$$

where $L_R = L_G\mathbb{E}[(1 + |X_0|)^\rho](1 + 2|R|)^{q-1} > 0$ and R is given by

$$R = \max\left\{ \left(\frac{2^{3(q-1)+1} L_G \mathbb{E}[(1 + |X_0|)^\rho]}{\eta} \right)^{\frac{1}{2r-1}}, \left(\frac{2^q L_G \mathbb{E}[(1 + |X_0|)^\rho]}{\eta} \right)^{\frac{1}{2r}} \right\}.$$

Under Assumption 1 and 2, $H(\theta, x)$ given in (5) is locally Lipschitz continuous in θ, which is explicitly stated in the following proposition. The proof follows the same idea in (Lovas et al., 2020, Proposition 2).

Proposition 5. *Let Assumption 1 and 2 hold. Then, one obtains that, for all* $\theta \in \mathbb{R}^d, x \in \mathbb{R}^m$,

$$|H(\theta, x) - H(\theta', x)| \leq L_H(1 + |x|)^\rho(1 + |\theta| + |\theta'|)^{2r+1}|\theta - \theta'|,$$

where $L_H = L_G + 8r\eta$.

Remark 6. *Let Assumption 1 and 2 hold. Then, Proposition 5 implies that* h *is locally Lipschitz continuous. That is, there exists a* $L_h > 0$ *such that for all* $\theta \in \mathbb{R}^d$,

$$|h(\theta) - h(\theta')| \leq L_h(1 + |\theta| + |\theta'|)^{2r+1}|\theta - \theta'|,$$

where $L_h = L_H(1 + \mathbb{E}[|X_0|)^\rho$.

The optimization problem (4) is closely linked to the problem of sampling from a target distribution $\pi_\beta(dz) \propto \exp(-\beta u(z))\mathrm{d}z$ with $\beta > 0$ since π_β concentrates around the minimizers of u for sufficiently large β. It is well-known that, under mild conditions (satisfied by Assumptions 1 and 2), the (overdamped) Langevin SDE given by

$$\mathrm{d}Z_t = -h(Z_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B_t, \tag{7}$$

for $t > 0$ where $\theta_0 \in \mathbb{R}^d$ is a (possibly random) initial condition, $\beta > 0$ is the so-called inverse temperature parameter, and $(B_t)_{t\geq 0}$ denoting a d-dimensional Brownian motion, admits $\pi_\beta(dz)$ as its unique invariant measure.

2.2 Mechanism of THεO POULA

We introduce the mechanism of TheoPouLa, which iterately updates as follows: for $n \in \mathbb{N}_0$ and $\theta_0^\lambda := \theta_0$,

$$\theta_{n+1}^\lambda := \theta_n^\lambda - \lambda H_{\lambda,c}(\theta_n^\lambda, X_{n+1}) + \sqrt{2\lambda\beta^{-1}}\xi_{n+1}, \tag{8}$$

where $\lambda > 0$ is the learning rate and $H_{\lambda,c} = \left(H_{\lambda,c}^{(1)}(\theta, x), \ldots, H_{\lambda,c}^{(d)}(\theta, x)\right) : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ is given by, for all $\theta \in \mathbb{R}^d, x \in \mathbb{R}^m$,

$$H_{\lambda,c}^{(i)}(\theta, x) = \underbrace{\frac{G^{(i)}(\theta, x)}{1 + \sqrt{\lambda}|G^{(i)}(\theta, x)|}}_{\text{taming function}} \underbrace{\left(1 + \frac{\sqrt{\lambda}}{\varepsilon + |G^{(i)}(\theta, x)|}\right)}_{\text{boosting function}} + \underbrace{\eta\frac{\theta^{(i)}|\theta|^{2r}}{1 + \sqrt{\lambda}|\theta|^{2r}}}_{\text{regularization term}}, \tag{9}$$

for $i = 1, \ldots, d$ and $0 < \varepsilon < 1$. In the formula of TheoPouLa (9), we call the functions $1 + \sqrt{\lambda}|G^{(i)}(\theta, x)|$ and $1 + \frac{\sqrt{\lambda}}{\varepsilon + |G^{(i)}(\theta,x)|}$ as the taming function and boosting function of the newly proposed algorithm, respectively.

TheoPouLa has several distinct features over the existing optimization methods in the literature. We give an intuitive explanation as to how these features are complementarily harmonized to efficiently tackle the exploding and vanishing gradient problems. We omit the regularization term, i.e., $\eta = 0$, and the noise term, $\sqrt{2\lambda\beta^{-1}}\xi_{n+1}$, throughout the exposition for simplicity. Also, we refer to λ as the *learning rate* and $|\Delta\theta_n^\lambda| := |\theta_{n+1}^\lambda - \theta_n^\lambda| = \frac{\lambda|G^{(i)}(\theta_n^\lambda, X_{n+1})|}{1+\sqrt{\lambda}|G^{(i)}(\theta_n^\lambda, X_{n+1})|} \times \left(1 + \frac{\sqrt{\lambda}}{\varepsilon + |G^{(i)}(\theta_n^\lambda, X_{n+1})|}\right)$ as the *stepsize* by the convention in Kingma and Ba (2015).

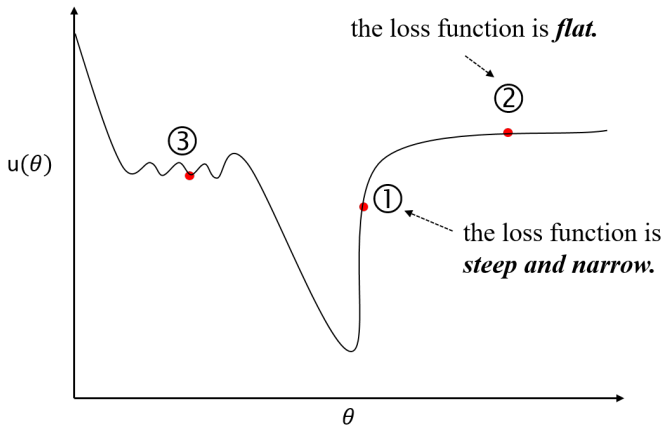


Figure 1: A landscape of a loss function.

Firstly, the new algorithm adopts the taming function to control the super-linearly growing gradient. In region ① of Figure 1 where the loss function is steep and narrow, i.e., the gradient is huge, it is ideal for the optimizer to take a small stepsize. This is effectively achieved because the growth of the taming function is proportional to G, which relieves the huge gradient. On the other hand, the boosting is close to one, i.e. it becomes useless in this case. The effectiveness of the taming function is confirmed in the motivating example in Section 3.1. In particular, we emphasize that the taming function of TheoPouLa is applied element-wise to scale the effective element-wise learning rate in contrast to TUSLA of Lovas et al. (2020). This significantly improves the performance of TheoPoula when solving high-dimensional optimization problems such as the training of neural network models.

Secondly, we first introduce the boosting function to accelerate training speed and prevent the vanishing gradient problem. When the current parameter is located in region ② of Figure 1 where the loss function is (almost) flat, i.e. the gradient is small, the boosting function helps TheoPoula to adaptively increase its stepsize. Specifically, the boosting function can increases the stepsize by up to $\sqrt{\lambda}/\varepsilon$, whereas the taming function barely contributes to the stepsize. We highlight that the taming and boosting functions do not interfere with each other in any adverse way. On the contrary, they complement each other in a harmonious way. We verify the effectiveness of the boosting function, which can be found in Section 3.4. The experiments show that the boosting function brings a significant improvement in test accuracy across different models and data sets for deep learning models.

Thirdly, TheoPouLa is quickly converted from adaptive learning rate methods to SGD. In the early training phase, TheoPouLa certainly behaves like adaptive learning rate methods. Then, when the current position is approaching an optimal solution, TheoPouLa is similar to SGD with a learning rate $(1 + \sqrt{\lambda}/\varepsilon)$. Consequently, TheoPouLa simultaneously attains two favorable features of fast training in adaptive learning rate methods and good generalization in SGD. The switching from adaptive learning rates to SGD has been also investigated by different strategies in Luo et al. (2019) and Keskar and Socher (2017).

Lastly, a scaled Gaussian noise, $\sqrt{2\lambda\beta^{-1}}\xi_{n+1}$, is added as a consequence of the discretization of the Langevin SDE. The term is essential to prove the convergence property of TheoPouLa. Moreover, adding properly scaled Gaussian noise allows the new algorithm

to escape from region ③ of Figure 1, local minima or saddle points, in a similar manner to the standard SGLD method, see Raginsky et al. (2017).

2.3 Convergence Analysis

In this subsection, we present the main convergence results of TheoPouLa in Wasserstein-1 and Wasserstein-2 distances which are defined in (3). The convergence is guaranteed when λ is less than $\lambda_{\max}$, which is given by

$$\lambda_{\max} = \min\left\{1, \frac{1}{4\eta^2}, \frac{1}{2^{14}\eta^2({}_{8l}\mathcal{C}_{4l})^2}\right\}. \tag{10}$$

where ${}_n\mathcal{C}_k$ is the binomial coefficient 'n choose k' and $l = 2r + 1$. Note that the learning rate restriction causes no issues as η is typically very small ($\eta \ll 1$).

Theorem 7 and Corollary 8 state the non-asymptotic estimates for the Wasserstein-1 and -2 distances between $\mathcal{L}\left(\theta_n^\lambda\right)$ and π_β. The proofs of the main results can be found in Section 4.

Theorem 7. *Let Assumption 1 and 2 hold. Then, for all $0 < \lambda \leq \lambda_{\max}$, $n \in \mathbb{N}_0$, we have that*

$$W_1\left(\mathcal{L}\left(\theta_n^\lambda\right), \pi_\beta\right) \leq C_1\sqrt{\lambda} + C_2 e^{-C_0\lambda n},$$

where C_0, C_1 and C_2 are explicitly given in Table 9. Moreover, the constants C_0, C_1, C_2 are independent of n and λ.

Corollary 8. *Let Assumption 1 and 2 hold. Then, for all $0 < \lambda \leq \lambda_{max}$, $n \in \mathbb{N}_0$, we have*

$$W_2\left(\mathcal{L}\left(\theta_n^\lambda\right), \pi_\beta\right) \leq C_3\lambda^{\frac{1}{4}} + C_4 e^{-C_5\lambda n},$$

where C_3, C_4 and C_5 are explicitly given in Table 9. Moreover, the constants C_3, C_4, C_5 are independent of n and λ.

We are now concerned with the expected excess risk of TheoPouLa, so called the optimization error of θ_n^λ, which is defined as

$$\mathbb{E}[u(\theta_n^\lambda)] - u(\theta^*), \tag{11}$$

where $\theta^* := \arg\min_{\theta\in\mathbb{R}^d} u(\theta)$. Using the result in Corollary 8, one can further obtain an error bound of the expected excess risk as stated in the below.

Theorem 9. *Let Assumption 1 and 2 hold. For any $n \in \mathbb{N}_0$, the expected excess risk of the n-th iterate of TheoPouLa is bounded by*

$$\mathbb{E}[u(\theta_n^\lambda)] - u(\theta^*) \leq C_6 W_2(\mathcal{L}(\theta_n^\lambda, \pi_\beta)) + \frac{\frac{d}{2}\log\left(\frac{Ke}{A}\left(\frac{B}{d}\beta + 1\right)\right) + \log 2}{\beta},$$

where $W_2(\mathcal{L}(\theta_n^\lambda), \pi_\beta)$ is given in Corollary 8 and constants A, B, K, C_6 are explicitly given in Table 9. Moreover, the constants A, B, K, C_6 are independent of n and λ.

Remark 10. *The constants C_0, C_1, C_2, C_3, C_4, C_6 are independent of n and λ, but might depend on β and d. In particular, the constants have exponential dependence on the dimension d because our nonconvex setting should encompass possible pathological scenarios. In particular, the exponential dependence on d only comes from the contraction property of the Langevin SDE in Lemma 17, inherited from the result in Eberle et al. (2019). In other words, if the contraction estimate can be improved under reasonable regularities, the exponential dependence on d is accordingly relaxed without affecting our analysis.*

Using Corollary 8, the expected excess risk of TheoPoula in Theorem 9 is rewritten as

$$\mathbb{E}[u(\theta_n^\lambda)] - u(\theta^*) \leq C_3 C_6 \lambda^{\frac{1}{4}} + C_4 C_6 e^{-C_5 \lambda n} + \frac{\frac{d}{2}\log\left(\frac{Ke}{A}\left(\frac{B}{d}\beta + 1\right)\right) + \log 2}{\beta}.$$

Then, the error bound of the expected excess risk in Theorem 9 can be interpreted via the following three steps: (i) For any $\delta > 0$, choose $\bar{\beta} > 0$ such that

$$\frac{\frac{d}{2}\log\left(\frac{Ke}{A}\left(\frac{B}{d}\bar{\beta} + 1\right)\right) + \log 2}{\bar{\beta}} \leq \frac{\delta}{3},$$

and fix $\bar{\beta}$, (ii) Then, pick and fix $\bar{\lambda} > 0$ such that

$$C_3 C_6 \bar{\lambda}^{\frac{1}{4}} \leq \frac{\delta}{3},$$

by using that C_3, C_6 are independent of λ, (iii) Lastly, choose $\bar{n} > 0$ such that

$$C_4 C_6 e^{-C_5 \bar{\lambda} \bar{n}} \leq \frac{\delta}{3}.$$

Therefore, for any $\delta > 0$, one can always find $(\bar{\lambda}, \bar{n}, \bar{\beta})$ that achieves the expected excess risk of TheoPoula being less than δ.

3. Numerical Experiments

This section provides extensive numerical experiments to demonstrate the empirical performance of TheoPouLa. In Section 3.1, we present a simple example to illustrate that popular stochastic optimization algorithms may fail to find the optimal solution in the presence of the super-linearly growing stochastic gradient. In Section 3.2, we present two real-world deep learning tasks such as image classification on CIFAR10 (Krizhevsky et al.) and CIFAR-100 (Krizhevsk 2009), and language modeling on Penn Treebank (Marcus et al. 1999). In Section 3.3, we investigate the effect of key hyperparameters λ, ε, β on the performance of TheoPouLa and the effectiveness of the boosting function.

3.1 Toy example

The super-linearly growing gradient and its effect on the performance of optimization methods are relatively under-studied because most relevant studies assume that the stochastic

gradient is global Lipscthiz continuous and bounded (Kingma and Ba 2015; Xu et al. 2018; Brosse et al. 2018; Duchi et al. 2011; Tieleman and Hinton 2012; Reddi et al. 2018; Chen et al. 2019; Liu et al. 2020; Luo et al. 2019; Zhuang et al. 2020). However, the assumptions are not true for the problem of training neural networks. This section provides a simple one-dimensional optimization problem that illustrates the convergence issue of popular optimization algorithms when the stochastic gradient is locally Lipschitz continuous that results in the super-linearly growing gradient[2]. Lovas et al. (2020) considers a similar example to show the stability of TUSLA using a different taming function.

Consider the following optimization problem:

$$\min_{\theta\in\mathbb{R}} u(\theta) = \min_{\theta\in\mathbb{R}} \mathbb{E}[U(\theta, X)], \tag{12}$$

where $U : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ is defined as

$$U(\theta,x) = \begin{cases} \theta^2\left(1+\mathbb{1}_{x\le 1}\right)+\theta^{30} & \text{if } |\theta|\le 1,\\ (2|\theta|-1)\left(1+\mathbb{1}_{x\le 1}\right)+\theta^{30} & \text{if } |\theta|>1,\end{cases}$$

and X is uniformly distributed over $(-2,2)$, that is, $f_X(x) = \frac{1}{4}\mathbb{1}_{|x|\le 2}$. Furthermore, the stochastic gradient $H : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ is given by

$$H(\theta,x) = \begin{cases} 2\theta\left(1+\mathbb{1}_{x\le 1}\right)+30\theta^{29} & \text{if } |\theta|\le 1,\\ 2(1+\mathbb{1}_{x\le 1})sgn(\theta)+30\theta^{29} & \text{if } |\theta|>1,\end{cases}$$

where $sgn(\cdot)$ is the sign function. Then, one can easily show that the stochastic gradient H is locally Lipschitz continuous:

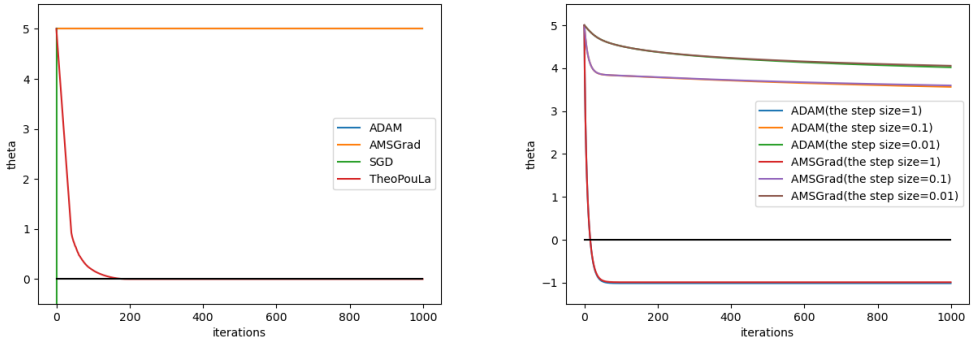
$$|H(\theta,x)-H(\theta',x)| \le 34(1+|\theta|+|\theta'|)^{28}|\theta-\theta'|,$$

for all $x\in\mathbb{R}$ and $\theta,\theta'\in\mathbb{R}$. Moreover, the minimum value of the optimization problem (12) is attained at $\theta = 0$.

We examine the behavior of SGD, Adam, and AMSGrad in solving the optimization problem (12) with the initial value $\theta_0 = 5$. For hyperparameters of the optimization algorithms, we use their default settings provided in PyTorch. More specifically, for Adam and AMSGrad, $\lambda = 0.001$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ are used. Figure 2(a) presents the trajectories of approximate solutions generated by the optimization algorithms, showing that SGD, Adam, and AMSGrad fail to converge to the optimal solution 0 even after $1,000$ iterations.

Intuitively, the undesirable phenomenon of Adam-type optimizers occurs because, in the update rule (2), the denominator $\sqrt{V}_n$, so-called the preconditioner, excessively dominates the numerator m_n, causing the vanishing gradient problem in the presence of the super-linealy growing gradient. On the other hand, SGD suffers from the exploding gradient problem due to the huge gradient. Moreover, Figure 2(b) highlights that the problematic behavior cannot be simply resolved by adjusting the learning rate within the Adam-type framework. On the contrary, TheoPouLa rapidly finds the optimal solution only after 200 iterations due to its taming function that controls the super-linearly growing gradient.

2. A function $f : \mathbb{R}^k\to\mathbb{R}^j$ for $k,j\in\mathbb{N}$ is said to be super-linearly growing if $\sup_{\theta\in\mathbb{R}^k}\frac{|f(\theta)|}{1+|\theta|} = \infty$.



(a) default settings (b) different step sizes

Figure 2: Performance of SGD, Adam, AMSGrad and TheoPouLa on an artificial example with the initial value $\theta_0 = 5.0$

3.2 Empirical performance on real data sets

We compare the performance of TheoPouLa with that of other popular optimization algorithms including Adam (Kingma and Ba 2015), AdaBelief (Zhuang et al. 2020), AdamP (Heo et al. 2021), AdaBound (Luo et al. 2019), AMSGrad (Reddi et al. 2018), RMSProp (Tieleman and Hinton 2012), SWATS (Keskar and Socher 2017), SGD (with momentum) and ASGD (Merity et al. 2018). In particular, we consider three deep convolution neural networks for image classification: VGG11 (Simonya and Zisserman 2015), ResNet34 (Ioffe and Szegedy 2016) and DenseNet121 (Huang et al. 2017) models. For language modeling, AWD LSTMs with 1, 2 and 3 layers are considered. Each experiment is run three times to compute the mean and standard deviation of the best accuracy on the test dataset.

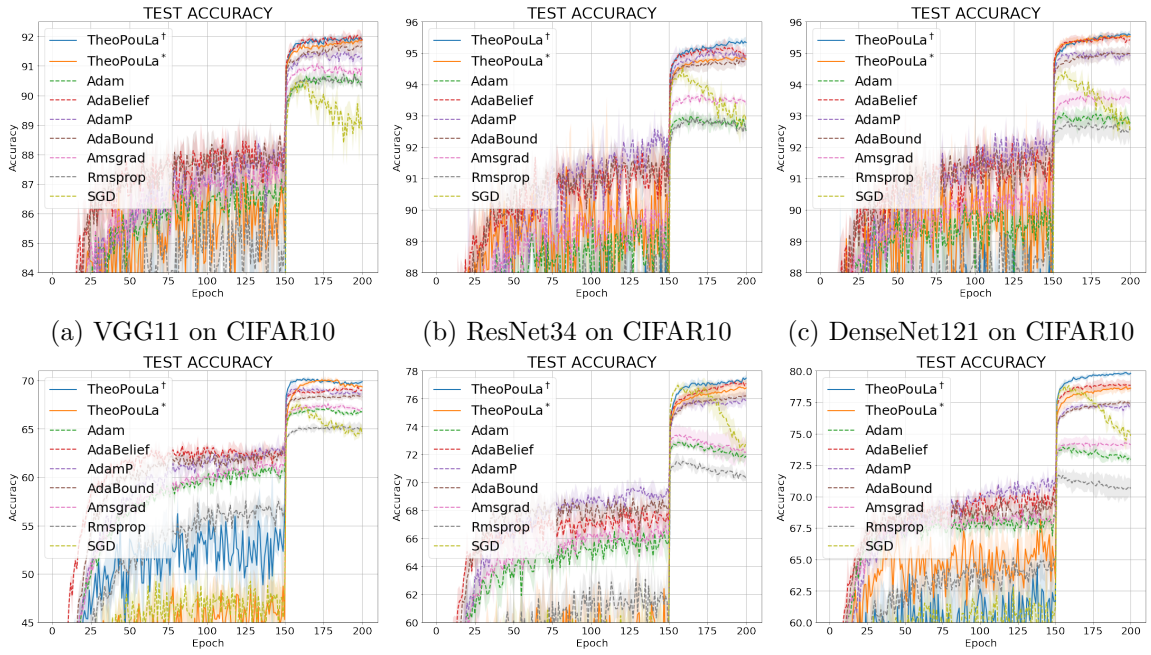
Recall that it is assumed that $r \geq \frac{q}{2} \geq \frac{1}{2}$ to obtain the main results. However, for the experiments in this section, we consider $r = 0$ in (4), which is equivalent to ℓ_2-regularization. This is justified by the fact that some form of dissipativity may already exist for specific problems such as the one considered here, although this has not been verified theoretical so far. In Section 3.4, we perform additional experiments with $r \geq \frac{1}{2}$, which show similar performance of THεO POULA as in Table 2 without any noticeable loss of accuracy. This demonstrates that there is no gap between theory and practice of our work.

Image classification. We replicate the experiments for image classification based on the official implementation of Zhuang et al. (2020) as it provides a reliable baseline of the experiments by comparing the performance of various optimization algorithms with extensive hyperparameter search. More specifically, all the models are trained for 500 epochs with batch size of 128. We set $\eta = 0.0005$ and $r = 0$ in (4), which is ℓ_2-regularization. We decay the initial learning rate by 10 after 150 epochs to all optimization algorithms.

For TheoPouLa, we search the optimal hyperparameters as follows: $\lambda \in \{1, 0.5, 0.1, 0.05, 0.01\}$, $\varepsilon \in \{1, 0.1, 0.01\}$ and $\beta \in \{10^8, 10^{10}, 10^{12}\}$. Regarding hyperparameter values of Adam, AdaBelief, AdamP, AdaBound, AMSGrad and RMSProp, the best hyperparameters are chosen among $\lambda \in \{1.0, 0.1, 0.01, 0.001\}$, $\beta_1 \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, $\beta_2 = 0.999$,

and $\epsilon = 10^{-8}$. For SGD, we set the momentum as 0.9 and search learning rate $\lambda \in \{10.0, 1.0, 0.1, 0.01, 0.001\}$[3].

Figure 3 shows test accuracy for VGG11, ResNet34 and DenseNet121 on CIFAR10 and CIFAR100. Table 2 shows the test accuracy for VGG11, ResNet34 and DenseNet121 on CIFAR10 and CIFAR100. As shown in Table 2, our algorithm achieves the highest accuracy and significantly outperforms the other optimization algorithms across all the experiments. In particular, TheoPouLa with the second best hyperparameter is even comparable to the AdaBelief (the state-of-the-art algorithm) and outperforms the other methods, validating that the solutions found by TheoPouLa yield good generalization performance. Also, the improvement of our algorithm is increasingly prominent as the models and datasets are more complicated and large-scale.



(a) VGG11 on CIFAR10 (b) ResNet34 on CIFAR10 (c) DenseNet121 on CIFAR10

(d) VGG11 on CIFAR100 (e) ResNet34 on CIFAR100 (f) DenseNet121 on CIFAR100

Figure 3: Test accuracy for VGG11, ResNet34 and DenseNet121 on CIFAR-10 and CIFAR-100. TheoPouLa† and THεOPOULA* represent the performances of TheoPouLa under the best and second best hyperparameters, respectively.

Language modeling. We conduct language modeling over the Penn Treebank (PTB) with AWD-LSTM models of Merity et al. (2018). It is reported that Non-monotonically Triggered ASGD (NT-ASGD) achieves state-of-the-art performance for this task. Motivated by this observation, we also consider averaged TheoPouLa, which is performed by averaging of trajectories of the parameters after a user-specified trigger Q, $\frac{1}{n-Q+1}\sum_{i=Q}^{n}\theta_i^\lambda$, instead of the last updated parameter θ_n^λ (Polyak and Juditsky 1992). Moreover, we use a

3. Our results for Adam, AdaBelief, AdaBound, AMSGrad, and RMSprop are consistent with the test accuracies reported in Luo et al. (2019) and Zhuang et al. (2020).

Table 2: Mean and standard deviation of the best accuracy for VGG11, ResNet34 and DenseNet121 on CIFAR10. TheoPouLa† and TheoPouLa* represent the performances of TheoPouLa with the best and second best hyperparameters, respectively. The numbers in parenthesises indicate the standard deviations.

dataset	CIFAR-10			CIFAR-100		
model	VGG	ResNet	DenseNet	VGG	ResNet	DenseNet
TheoPouLa†	**92.30** (0.055)	**95.43** (0.095)	**95.66** (0.066)	**70.31** (0.117)	**77.60** (0.208)	**79.90** (0.133)
TheoPouLa*	91.92 (0.119)	94.92 (0.076)	95.59 (0.067)	70.24 (0.227)	76.88 (0.536)	78.76 (0.269)
AdaBelief (baseline)	92.17 (0.035)	95.29 (0.196)	95.58 (0.095)	69.50 (0.111)	77.33 (0.172)	79.12 (0.382)
Adam	90.79 (0.075)	93.11 (0.184)	93.21 (0.240)	67.30 (0.137)	73.02 (0.231)	74.03 (0.334)
AdamP	91.68 (0.162)	95.18 (0.116)	95.17 (0.079)	69.41 (0.297)	76.14 (0.347)	77.58 (0.091)
AdaBound	91.81 (0.272)	94.83 (0.131)	95.05 (0.176)	68.61 (0.312)	76.27 (0.256)	77.56 (0.120)
AMSGrad	91.24 (0.115)	93.76 (0.108)	93.74 (0.236)	67.71 (0.291)	73.51 (0.692)	74.50 (0.416)
RMSProp	90.82 (0.201)	93.06 (0.120)	92.89 (0.310)	65.45 (0.394)	71.79 (0.287)	71.75 (0.632)
SGD	90.73 (0.090)	94.61 (0.280)	94.46 (0.159)	67.78 (0.320)	77.16 (0.214)	78.95 (0.312)
SWATS	87.29 (4.210)	94.76 (0.565)	95.04 (0.339)	N/A	73.86 (3.928)	78.81 (1.812)

trigger strategy which starts the averaging when no improvement in the validation metric is seen for a patience number of epochs. For our experiments, we set the patience number to 5. Due to a limited computation budget, we only test NT-ASGD, AdaBelief, and TheoPouLa rather than investigating the various optimization algorithms used in the image classification. Since AdaBelief significantly outperforms the other optimization algorithms including vanilla SGD, AdaBound, Yogi (Zaheer et al. 2018), Adam, MSVAG (Balles and Hennig 2018), RAdam, Fromage and AdamW (Loshchilov and Hutter 2019) in the same experiment, we believe that it is enough to compare the performance of AdaBelief, NT-ASGD and TheoPouLa.

For a fair comparison, the averaging scheme has also been applied to AdaBelief. However, it turns out that the averaging scheme does not improve the performance of AdaBelief. Instead, AdaBelief uses a development-based learning rate decay, which decreases the learning rate by a constant factor δ if the model does not attain a new best value for k epochs. We search the optimal learning rate schedule among $\delta \in \{0.1, 0.5\}$ and $k \in \{5, 10, 20\}$. For ASGD and TheoPouLa, a constant learning rate is used without a learning rate decay. Moreover, in order to compare with the baseline, we apply gradient clipping of 0.25 to all three optimization algorithms.

We train the AWD LSTM with 1,2 and 3 layers for 750 epochs with 20 batch size. The details of models can be found in the official implementation of AWD-LSTM [4]. For NT-

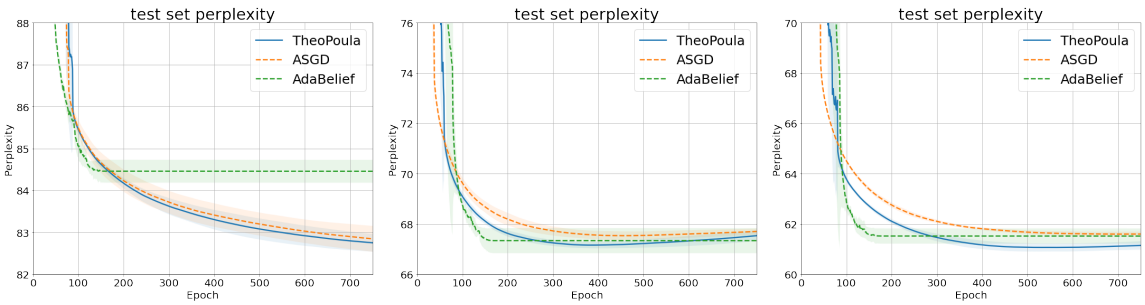
4. https://github.com/salesforce/awd-lstm-lm

Table 3: Test perplexity for language modeling tasks on PTB. Lower is better.

# of layers	1-layer	2-layer	3-layer
TheoPouLa	**82.75** (0.209)	**67.15** (0.126)	**61.07** (0.161)
ASGD (baseline)	82.85 (0.308)	67.53 (0.171)	61.60 (0.094)
AdaBelief	84.46 (0.272)	67.34 (0.496)	61.52 (0.302)

ASGD and averaged TheoPouLa, the constant learning rate of 30 is used for 2 and 3-layer LSTMs. For 1-layer LSTMs, we set λ to 10. Moreover, we fix the following hyperparameters: $\varepsilon = 100$ and $\beta = 10^{10}$ across all the experiments. For AdaBelief, we used the best hyperparameters reported in Zhuang et al. (2020). That is, we use $\lambda = 0.01$ and $\epsilon = 10^{-12}$ for 2 and 3-layer LSTMs, and $\lambda = 0.001$ and $\epsilon = 10^{-16}$ for 1-layer LSTMs where $\beta_1 = 0.9$ and $\beta_2 = 0.999$ are fixed.

The performance of each optimization algorithm is measured by test perplexity (The lower is better). Figure 4 displays test perplexity of different algorithms for different AWD-LSTM models on PTB. Table 3 shows that TheoPouLa attains the lower test perplexity against the baselines for AWD-LSTM with 1, 2, and 3-layers. AdaBelief shows a comparable performance with ASGD for 2-layer and 3-layer models.



(a) 1-layer (b) 2-layer (c) 3-layer

Figure 4: Test perplexity for 1, 2 and 3-layer AWD-LSTMs on PTB

3.3 Effect of λ, ε, and β on the performance of TheoPouLa

This section perform a sensitive analysis to understand the effect of key hyperparameters λ, ε, and β, on the performance of TheoPouLa. We consider image classification for VGG11 and ResNet34 on CIFAR10 and CIFAR100. As in Section 3.2, we train the models for 200 epochs with 128 batch size and then evaluate their test accuracy under varying λ, ε, and β.

We first report the effect of λ on the performance of TheoPouLa. To see this, we evaluate the models trained with $\lambda \in \{0.5, 0.1, 0.05, 0.01\}$, $\varepsilon = 0.1$, $\beta = 10^{10}$, $r = 0$, and $\eta = 0.0005$. As shown in Table 4, $\lambda = 0.1$ yields the best accuracy for VGG on CIFAR10 and CIFAR100, and ResNet on CIFAR10, but $\lambda = 0.05$ achieves the highest accuracy for ResNet on CIFAR100.

Table 4: The test accuracy for VGG11 and ResNet34 on CIFAR-10 and CIFAR-100 with different λ. We report mean and standard deviation of the accuracy from three repetitive experiments.

		λ			
model	dataset	0.5	0.1	0.05	0.01
VGG	CIFAR10	42.31 (4.996)	92.14 (0.201)	91.82 (0.081)	89.89 (0.060)
	CIFAR100	31.59 (2.929)	70.57 (0.279)	68.96 (0.205)	63.98 (0.185)
ResNet	CIFAR10	91.23 (0.312)	95.41 (0.175)	95.09 (0.117)	93.54 (0.277)
	CIFAR100	1.90 (0.316)	74.02 (0.609)	77.31 (0.248)	75.49 (0.066)

The hyperparameter ε controls the intensity of the boosting function as the stepsize can increase by up to $\sqrt{\lambda}/\varepsilon$. The effect of the boosting function can be exaggerated when ε is small, whereas a large ε depresses the role of the boosting function. To see the effect of ε, we fix $\lambda = 0.1$ and $\beta = 10^{10}$ and then conduct experiments with different $\varepsilon \in \{1, 0.1, 0.01, 0.001\}$. Table 5 summarizes the test accuracy with varying ε. TheoPouLa with $\varepsilon = 0.1$ shows the highest accuracy for VGG on CIFAR10 and CIAR100, and ResNet on CIFAR10. On the other hand, $\varepsilon = 0.01$ is the best hyperparameter for ResNet on CIFAR100. Moreover, it is observed that too small ε, 0.001, leads to unstability for VGG on CIFAR10.

Table 5: The test accuracy for VGG11 and ResNet34 on CIFAR-10 and CIFAR-100 with different ε. We report mean and standard deviation of the accuracy from three repetitive experiments.

		ε			
model	dataset	1	0.1	0.01	0.001
VGG	CIFAR10	91.80 (0.262)	92.14 (0.201)	86.75 (0.489)	26.03 (14.473)
	CIFAR100	68.75 (0.523)	70.57 (0.279)	60.55 (0.202)	44.74 (0.383)
ResNet	CIFAR10	91.23 (0.312)	95.41 (0.175)	95.09 (0.117)	93.54 (0.277)
	CIFAR100	1.90 (0.316)	74.02 (0.609)	77.31 (0.248)	75.49 (0.066)

Lastly, we evaluate the effect of the inverse temperature $\beta > 0$, which is a unique feature of Langevin based algorithms. Intuitively, a small inverse temperature offers relatively strong random shocks at each iteration, which is helpful to escape sharp local minima, so called *the exploration effect*. On the other hand, the solutions generated with a large inverse temperature explores the local geometry of the objective function, so called *the exploitation effect*. To leverage the trade-off of β, it is desirable to apply simulated annealing (Mangoubi and Vishnoi 2018) and simulated tempering (Lee et al. 2018), which often requires intensive effort for the hyperparameter tuning. In this paper, we fix β as a constant during the training. We obtain the test accuracy of TheoPouLa with different values of β ranging

from $\{10^4, 10^6, 10^8, 10^{10}, 10^{12}\}$. The other hyperparameters are as follows: $\lambda = 0.1$, $\varepsilon = 0.1$ for VGG on CIFAR10 and CIFAR100, and ResNet on CIFAR10, $\lambda = 0.05$, $\varepsilon = 0.01$ for ResNet on CIFAR100. Table 6 shows that TheoPouLa achieves its best performance when β is large, say $10^8 \sim 10^{12}$. We note that the result in Table 6 is consistent with the cold posterior effect implying that a large inverse temperature β improves the model's predictive power, see Aitchison (2021) and Wenzel et al. (2020).

Table 6: The accuracy for VGG11 and ResNet34 on CIFAR-10 and CIFAR-100 with different β. We report mean and standard deviation of the accuracy from three repetitive experiments.

		β				
model	dataset	10^4	10^6	10^8	10^{10}	10^{12}
VGG	CIFAR10	73.10 (0.407)	91.53 (0.141)	92.31 (0.055)	92.29 (0.120)	92.10 (0.023)
	CIFAR100	20.69 (0.718)	70.0 (0.343)	70.28 (0.124)	70.16 (0.110)	70.31 (0.117)
ResNet	CIFAR10	80.84 (0.264)	94.67 (0.145)	95.42 (0.117)	95.34 (0.141)	95.43 (0.095)
	CIFAR100	63.58 (0.103)	77.22 (0.291)	77.4 (0.036)	77.6 (0.208)	77.53 (0.143)

3.4 Additional Experiments

Experiments with $r \geq \frac{1}{2}$. In Section 3.2, we set $r = 0$ in (4) as ℓ_2-regularization is widely used in practice. However, our main results are derived with $r \geq \frac{q}{2} \geq \frac{1}{2}$ to theoretically ensure the dissipativity condition. In this section, we perform additional experiments with $r \geq \frac{1}{2}$ to demonstrate that there is no gap between theory and practice.

When the regularization parameter r is sufficiently large and the dimension d is big, $|\theta|^{2r}$ becomes extremely large. As a result, the stochastic gradient of the regularization term $\eta \frac{\theta^{(i)}|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}}$ in (9) approximately ends up with $\frac{\eta}{\sqrt{\lambda}}\theta^{(i)}$, which is equivalent to the same regularization effect of ℓ_2-regularization. Therefore, by choosing $\eta = 5 \times 10^{-4}\sqrt{\lambda}$ and large r, the performance of the models with $\eta = 5 \times 10^{-4}\sqrt{\lambda}$ and large r is similar to that of the models with ℓ_2-regularization (i.e., $r = 0$) and $\eta = 5 \times 10^{-4}$.

Table 7 shows that the accuracy for VGG, ResNet and DenseNet on CIFAR10 and CIFAR100 with $r = 10$ and $\eta = 5 \times 10^{-4}\sqrt{\lambda}$. We use the same best hyperparameters used in Section 3.2. As shown in Table 7, one observes the experiments with large r is highly similar to the results in Section 3.2.

Table 7: The accuracy for VGG11, ResNet34 and DenseNet121 on CIFAR10 and CIFAR100 with $r = 10$.

dataset	CIFAR10			CIFAR100		
model	VGG	ResNet	DenseNet	VGG	ResNet	DenseNet
TheoPouLa	92.2	95.38	95.69	70.07	77.78	80.47

Effectiveness of the boosting function. This subsection empirically tests the effectiveness of the boosting function in our algorithm. TheoPouLa without the boosting function is given by

$$H_{\lambda,c}^{(i)}(\theta,x) = \frac{G^{(i)}(\theta,x)}{1+\sqrt{\lambda}|G^{(i)}(\theta,x)|} + \eta\frac{\theta^{(i)}|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}}. \tag{13}$$

Indeed, this is a special case of TheoPouLa with $\varepsilon = \infty$. We train VGG11, ResNet34 and DenseNet121 on CIFAR-10 and CIFAR-100 using TheoPoula without the boosting function. We use the hyperparameters as the best hyperparameters of TheoPouLa in Section 3.2. Table 8 clearly shows that the performance of TheoPouLa without the boosting function deteriorates, confirming that the addition of the boosting function brings meaningful increase in test accuracy.

Table 8: The best accuracy for VGG11, ResNet34 and DenseNet121 on CIFAR-10 and CIFAR-100 obtained from TheoPouLa with/without the boosting function.

dataset	CIFAR-10			CIFAR-100		
model	VGG	ResNet	DenseNet	VGG	ResNet	DenseNet
TheoPouLa	**92.30**	**95.43**	**95.66**	**70.31**	**77.60**	**79.90**
TheoPouLa ($\varepsilon = \infty$)	91.48	94.31	94.22	68.11	75.91	77.99

4. Overview of the Proofs

This section provides an overview of the proofs of Theorem 7, Corollary 8 and Theorem 9. In Section 4.1, we introduce suitable Lyapunov functions and auxiliary processes, which are necessary to analyze the convergence of TheoPouLa. Then, necessary momentum bound for the auxiliary processes are estimated. Lastly, the proofs of main results can be found in Section 4.3.

4.1 Auxiliary processes.

For each $p \geq 1$, define the Lyapunov function V_p by, for all $\theta \in \mathbb{R}^d$,

$$V_p(\theta) := (1+|\theta|^2)^{\frac{p}{2}}, \tag{14}$$

and similarly, define $\mathrm{v}_p(x) = (1+x^2)^{\frac{p}{2}}$ for $x \geq 0$. Both functions are continuously differentiable and $\lim_{|\theta|\to\infty} \nabla V_p(\theta)/V_p(\theta) = 0$. Also, denote by $Z_t^\lambda := Z_{\lambda t}, t \in \mathbb{R}_+$, the time-changed Langevin dynamics of (7) given by

$$\mathrm{d}Z_t^\lambda = -\lambda h(Z_t^\lambda)\mathrm{d}t + \sqrt{2\lambda\beta^{-1}}\mathrm{d}B_t^\lambda, \tag{15}$$

with $t > 0$, $Z_0 := \theta_0$, where $B_t^\lambda := B_{\lambda t}/\sqrt{\lambda}$ is a d-dimensional standard Brownian motion. We then consider the continuous-time interpolation of the TheoPouLa algorithm (8), denoted by $(\bar{\theta}_t^\lambda)_{t\in\mathbb{R}_+}$ as

$$\mathrm{d}\bar{\theta}_t^\lambda = -\lambda H_\lambda\left(\bar{\theta}_{\lfloor t\rfloor}^\lambda, X_{\lceil t\rceil}\right)\mathrm{d}t + \sqrt{2\lambda\beta^{-1}}\mathrm{d}B_t^\lambda \tag{16}$$

with the initial condition $\bar{\theta}_0^\lambda = \theta_0$. Here, $\lfloor x \rfloor$ denotes the integer part of a positive real x and $\lceil x \rceil := \lfloor x \rfloor + 1$. Due to the construction of (16), the law of interpolated process is equivalent to the law of the TheoPouLa algorithm at grid points, i.e., $\mathcal{L}(\bar{\theta}_n^\lambda) = \mathcal{L}(\theta_n^\lambda)$, for all $n \in \mathbb{N}_0$.

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

$$\mathrm{d}\zeta_t^{s,v,\lambda} = -\lambda h\left(\zeta_t^{s,v,\lambda}\right)\mathrm{d}t + \sqrt{2\lambda\beta^{-1}}\mathrm{d}B_t^\lambda, \tag{17}$$

with the initial condition $\zeta_s^{s,v,\lambda} := v \in \mathbb{R}^d$.

Definition 11. *For each fixed $\lambda > 0$ and $n \in \mathbb{N}_0$, define $\bar{\zeta}_t^{\lambda,n} := \zeta_t^{nT,\bar{\theta}_{nT}^\lambda,\lambda}, t \geq nT$ where $T = \lfloor 1/\lambda \rfloor$ and $\zeta_t^{s,v,\lambda}$ is given in* (17).

4.2 Primary estimates

We provide moment estimates for $(\theta_n^\lambda)_{n\geq 1}$ in the following two lemmas. Recall that $\lambda_{\max}$ is defined in (10).

Lemma 12. *Let Assumption 1 and 2 hold. Then there exists $M_0 > 0$ such that ,for $0 < \lambda \leq \lambda_{\max}$, $n \in \mathbb{N}_0$,*

$$\mathbb{E}|\theta_{n+1}^\lambda|^2 \leq \left(1 - \frac{\eta}{2}\sqrt{\lambda}\right)^n \mathbb{E}|\theta_0|^2 + \left[5M_0^2 + \frac{4d}{\eta}\left(\beta^{-1} + 4\right) + \frac{8\sqrt{d}M_0}{\eta} + 4\eta M_0^2\right],$$

and

$$\sup_n \mathbb{E}|\theta_{n+1}^\lambda|^2 \leq \mathbb{E}|\theta_0|^2 + \left[5M_0^2 + \frac{4d}{\eta}\left(\beta^{-1} + 4\right) + \frac{8\sqrt{d}M_0}{\eta} + 4\eta M_0^2\right].$$

Proof. See Appendix B. □

Lemma 13. *Let Assumption 1 and 2 hold. Then, one obtains that, for all $0 < \lambda < \lambda_{\max}$), $n \in \mathbb{N}_0$, $p \in [1, 8(2r+1)]$,*

$$\mathbb{E}|\theta_{n+1}^\lambda|^{2p} \leq (1 - \eta^2\lambda)^n \mathbb{E}|\theta_0^\lambda|^{2p} + \frac{A_p}{\eta^2},$$

and

$$\sup_{n\in\mathbb{N}} \mathbb{E}|\theta_{n+1}^\lambda|^{2p} \leq \mathbb{E}|\theta_0^\lambda|^{2p} + \frac{A_p}{\eta^2},$$

where A_p is given in Table 9.

Proof. See Appendix B. □

Using Lemma 12 and Lemma 13, one can establish the fourth moment bounds for $(\bar{\theta}_t^\lambda)_{t\geq nT}$ and $(\bar{\zeta}_t^{\lambda,n})_{t\geq nT}$. To this end, we introduce a drift condition for the Lyapunov function, which is stated in the following lemma.

Lemma 14. *Let Assumption 1 and 2 hold. Then, one obtains, for any $p \in [2, \infty) \cap \mathbb{N}$, $\theta \in \mathbb{R}^d$,*

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

where $\bar{c}(p) = Ap/4$, $\tilde{c}(p) = (3/4)Ap\mathrm{v}_p(\bar{M}_p)$, $\bar{M}_p = (1/3 + 4B/(3A) + 4d/(3A\beta) + 4(p-2)/(3A\beta))^{1/2}$, and A, B are explicitly given in Remark 3.

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

Lemma 15. *Let Assumption 1 and 2 hold. Then, one obtains, for $n \in \mathbb{N}_0$, $0 < \lambda \leq \lambda_{\max}$,*

$$\mathbb{E}[V_4(\bar{\theta}^{\lambda}_{nT})] \leq 2\mathbb{E}|\theta_0|^4 + 2 + 2\frac{A_2}{\eta^2},$$

where A_2 is given in Table 9.

Proof. From the definition of the Lyapunov function V_m given in (14) and Lemma 13, we have

$$\begin{aligned} \mathbb{E}[V_4(\bar{\theta}^{\lambda}_{nT})] &= \mathbb{E}[(1 + |\bar{\theta}^{\lambda}_{nT}|^2)^2] \\ &\leq 2 + 2\mathbb{E}|\bar{\theta}^{\lambda}_{nT}|^4 \\ &\leq 2 + 2\mathbb{E}|\theta_0|^4 + 2\frac{A_2}{\eta^2}. \end{aligned}$$

□

Lemma 16. *Let Assumption 1 and 2 hold. Then, one obtains that, for $n \in \mathbb{N}_0$, $0 < \lambda \leq \lambda_{\max}$, $t \in (nT, (n+1)T]$,*

$$\mathbb{E}[V_4(\bar{\zeta}^{\lambda,n}_t)] \leq 2\mathbb{E}|\theta_0|^4 + 2 + \frac{2A_2}{\eta^2} + \frac{\tilde{c}(4)}{\bar{c}(4)},$$

where $\bar{c}(4)$, $\tilde{c}(4)$, A_2 are given in Table 9.

Proof. See Appendix B. □

Denote by $\mathcal{P}_{V_2}(\mathbb{R}^d)$ the subset of $\mathcal{P}(\mathbb{R}^d)$ such that $\mu \in \mathcal{P}_{V_2}(\mathbb{R}^d)$ satisfies $\int_{\mathbb{R}^d} V_2(\theta)\mu(d\theta) < \infty$. Moreover, we consider the following functional, for all $p \geq 1$, $\mu, \nu \in \mathcal{P}_{V_2}(\mathbb{R}^d)$,

$$w_{1,2}(\mu,\nu) := \inf_{\zeta \in \mathcal{C}(\mu,\nu)} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \left[1 \wedge |\theta - \theta|'\right] \left[\left(1 + V_2(\theta) + V_2\left(\theta'\right)\right)\zeta\left(\mathrm{d}\theta\mathrm{d}\theta'\right)\right., \tag{18}$$

where $\mathcal{C}(\mu,\nu)$ is defined in (3). For any $\mu, \nu \in \mathcal{P}_{V_2}(\mathbb{R}^d)$, the following inequalities hold:

$$W_1(\mu,\nu) \leq w_{1,2}(\mu,\nu), \quad W_2(\mu,\nu) \leq \sqrt{2w_{1,2}(\mu,\nu)}. \tag{19}$$

The following lemma states the contraction property of the Langevin SDE (15) in $w_{1,2}$, which is a key result of our analysis.

Lemma 17. *Let Assumption 1 and 2 hold. Let $Z'_t, t \in \mathbb{R}_+$ be the solution of the Langevin SDE (7) with initial condition $Z'_0 = \theta'_0$ which is independent of $\mathcal{G}_\infty$ and $|\theta'_0| \in L^2$. Then, one obtains*

$$w_{1,2}\left(\mathcal{L}(Z_t^\lambda), \mathcal{L}(Z'_t)\right) \leq \hat{c}e^{-C_0 t} w_{1,2}\left(\mathcal{L}\left(\theta_0\right), \mathcal{L}\left(\theta'_0\right)\right)$$

where $w_{1,2}$ is given in (18). The constant C_0 is given by

$$C_0 := \min\{\bar{\phi}, \bar{c}(2), 4\tilde{c}(2)\bar{c}(2)\epsilon\}/2,$$

where $\bar{c}(2) = A/2$, $\tilde{c}(2) = (3/2)A\mathrm{v}_2(\bar{M}_2)$, $\bar{M}_2 = (1/3+4B/(3A)+4d/(3A\beta))^{1/2}$, $\bar{\phi}$ is defined by

$$\bar{\phi} = \left(\sqrt{8\beta\pi/L_R}\dot{c}_0 \exp\left\{\left(\dot{c}_0\sqrt{\beta L_R/8} + \sqrt{8/(\beta L_R)}\right)^2\right\}\right)^{-1},$$

where L_R is defined in Proposition 4, and ϵ is chosen such that the following inequality is satisfied

$$\epsilon \leq 1 \wedge \left(4\tilde{c}(2)\sqrt{2\beta\pi/L_R}\int_0^{\dot{c}_1} \exp\left\{\left(s\sqrt{\beta L_R/8} + \sqrt{8/(\beta L_R)}\right)^2\right\} \mathrm{d}s\right)^{-1},$$

with $\dot{c}_0 = 2\left(4\tilde{c}(2)(1+\bar{c}(2))/\bar{c}(2) - 1\right)^{1/2}$ and $\dot{c}_1 = 2\left(\tilde{c}(2)/\bar{c}(2) - 1\right)^{1/2}$. Moreover, the constant $\hat{c}$ is given by

$$\hat{c} = 2(1+\dot{c}_0)\exp(\beta L_R \dot{c}_0^2/8 + 2\dot{c}_0)/\epsilon.$$

Proof. See (Chau et al., 2019, Proposition 3.14). □

We are now able to provide non-asymptotic bounds in the Wasserstein distances between the Auxiliary processes, namely $W_2\left(\mathcal{L}(\bar{\theta}_t^\lambda), \mathcal{L}(\bar{\zeta}_t^{\lambda,m})\right)$, $W_1\left(\mathcal{L}(\bar{\zeta}_t^{\lambda,m}), \mathcal{L}(Z_t^\lambda)\right)$, and $W_2\left(\mathcal{L}(\bar{\zeta}_t^{\lambda,m}), \mathcal{L}(Z_t^\lambda)\right)$. These results are key components to derive our main results.

Lemma 18. *Let Assumptions 1 and 2 hold. Then, one obtains, for all $0 < \lambda \leq \lambda_{max}$, $n \in \mathbb{N}_0$, $t \in (nT, (n+1)T]$,*

$$W_2\left(\mathcal{L}(\bar{\theta}_t^\lambda), \mathcal{L}(\bar{\zeta}_t^{\lambda,n})\right) \leq \sqrt{\lambda}\sqrt{e^{3L_R}(\bar{C}_1 + \bar{C}_2 + \bar{C}_3)}$$

where the constants $\bar{C}_1$, $\bar{C}_2$, $\bar{C}_3$ are given explicitly in Table 9, and L_R is given in Proposition 4.

Proof. See Appendix B. □

Lemma 19. *Let Assumptions 1 and 2 hold. Then, one obtains, for all $0 < \lambda \leq \lambda_{max}$, $n \in \mathbb{N}_0$, $t \in (nT, (n+1)T]$,*

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

where z_1 is given explicitly in Table 9.

Proof. See Appendix B. □

Lemma 20. *Let Assumptions 1 and 2 hold. Then, one obtains, for all* $0 < \lambda \leq \lambda_{max}$, $n \in \mathbb{N}_0$, $t \in (nT, (n+1)T]$,

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

where z_2 *is given explicitly in Table 9.*

Proof. See Appendix B. □

4.3 Proofs of main results

To derive non-asymptotic upper bounds for $W_1\left(\mathcal{L}(\theta_t^{\lambda}), \pi_\beta\right)$ and $W_2\left(\mathcal{L}(\theta_t^{\lambda}), \pi_\beta\right)$, we consider the following decomposition in terms of the auxiliary processes $\bar{\theta}_t^{\lambda}$, $\bar{\zeta}_t^{\lambda,n}$, and Z_t^{λ} as follows:

$$W_j\left(\mathcal{L}(\theta_t^{\lambda}), \pi_\beta\right) \leq W_j\left(\mathcal{L}(\bar{\theta}_t^{\lambda}), \mathcal{L}(\bar{\zeta}_t^{\lambda,n})\right) + W_j\left(\mathcal{L}(\bar{\zeta}_t^{\lambda,n}), \mathcal{L}\left(Z_t^{\lambda}\right)\right) + W_j\left(\mathcal{L}(Z_t^{\lambda}), \pi_\beta\right),$$

for $j = 1, 2$.

Proof of Theorem 7. Observe that $W_1\left(\mathcal{L}(\theta_n^{\lambda}), \mathcal{L}(Z_t^{\lambda})\right)$ is decomposed as follows: for all $t \in (nT, (n+1)T]$, $n \in \mathbb{N}_0$, $0 < \lambda \leq \lambda_{\max}$,

$$W_1\left(\mathcal{L}(\bar{\theta}_t^{\lambda}), \pi_\beta\right) \leq W_1\left(\mathcal{L}(\bar{\theta}_t^{\lambda}), \mathcal{L}(Z_t^{\lambda})\right) + W_1\left(\mathcal{L}(Z_t^{\lambda}), \pi_\beta\right). \tag{20}$$

Then, from the results of Lemma 18 and 19, the first term in (20) is estimated by, for $t \in (nT, (n+1)T]$, $0 < \lambda \leq \lambda_{\max}$, $n \in \mathbb{N}_0$,

$$\begin{aligned}
W_1\left(\mathcal{L}(\bar{\theta}_t^{\lambda}), \mathcal{L}(Z_t^{\lambda})\right) &\leq W_1\left(\mathcal{L}(\bar{\theta}_t^{\lambda}), \mathcal{L}(\bar{\zeta}_t^{\lambda,n})\right) + W_1\left(\mathcal{L}(\bar{\zeta}_t^{\lambda,n}), \mathcal{L}(Z_t^{\lambda})\right) \\
&\leq W_2\left(\mathcal{L}(\bar{\theta}_t^{\lambda}), \mathcal{L}(\bar{\zeta}_t^{\lambda,n})\right) + W_1\left(\mathcal{L}(\bar{\zeta}_t^{\lambda,n}), \mathcal{L}(Z_t^{\lambda})\right) \\
&\leq \sqrt{\lambda}(\sqrt{e^{3L_R}(\bar{C}_1 + \bar{C}_2 + \bar{C}_3)} + z_1) \\
&\leq C_1\sqrt{\lambda}, \qquad (21)
\end{aligned}$$

where $C_1 := \sqrt{e^{3L_R}(\bar{C}_1 + \bar{C}_2 + \bar{C}_3)} + z_1$.

Consequently, using (21), (19), and Lemma 17, we derive, for $t \in (nT, (n+1)T]$, $0 < \lambda \leq \lambda_{\max}$, $n \in \mathbb{N}_0$,

$$\begin{aligned}
W_1\left(\mathcal{L}(\bar{\theta}_t^{\lambda}), \pi_\beta\right) &\leq \sqrt{\lambda}(\sqrt{e^{3L_R}(\bar{C}_1 + \bar{C}_2 + \bar{C}_3)} + z_1) + w_{1,2}\left(\mathcal{L}(Z_t^{\lambda}), \pi_\beta\right) \\
&\leq \sqrt{\lambda}(\sqrt{e^{3L_R}(\bar{C}_1 + \bar{C}_2 + \bar{C}_3)} + z_1) + \hat{c}e^{-C_0\lambda t} w_{1,2}(\theta_0, \pi_\beta) \\
&\leq \sqrt{\lambda}(\sqrt{e^{3L_R}(\bar{C}_1 + \bar{C}_2 + \bar{C}_3)} + z_1) \\
&+ \hat{c}e^{-C_0\lambda t}\left[1 + \mathbb{E}[V_2(\theta_0)] + \int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(d\theta)\right] \\
&\leq C_1\sqrt{\lambda} + C_2 e^{-C_0 n},
\end{aligned}$$

where $C_2 := \hat{c}\Big(1 + \mathbb{E}[V_2(\theta_0)] + \int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(d\theta)\Big)$, which yields that, for all $n \in \mathbb{N}_0$,

$$W_1\left(\mathcal{L}(\bar{\theta}^\lambda_{nT}), \pi_\beta\right) \le C_1\sqrt{\lambda} + C_2 e^{-C_0 n}.$$

By setting nT to n and noticing that $n\lambda \le n/T$ in the above inequality, we obtain the desired result. □

Proof of Corollary 8. Similar to the proof of Theorem 7, we consider the following decomposition: for $t \in (nT, (n+1)T]$, $0 < \lambda \le \lambda_{\max}$, $n \in \mathbb{N}_0$,

$$W_2\left(\mathcal{L}(\theta^\lambda_t), \pi_\beta\right) \le W_2\left(\mathcal{L}(\bar{\theta}^\lambda_t), \mathcal{L}(\bar{\zeta}^{\lambda,n}_n)\right) + W_2\left(\mathcal{L}(\bar{\zeta}^{\lambda,n}_t), \mathcal{L}(Z^\lambda_t)\right) + W_2\left(\mathcal{L}(Z^\lambda_t), \pi_\beta\right).$$

Then, using Lemma 18, Lemma 20 and (19), one further obtains that

$$W_2\left(\mathcal{L}(\theta^\lambda_t), \pi_\beta\right) \le \sqrt{e^{3L_R}(\bar{C}_1 + \bar{C}_2 + \bar{C}_3)}\sqrt{\lambda} + z_2\lambda^{\frac{1}{4}} + \sqrt{2w_{1,2}(\mathcal{L}(Z^\lambda_t), \pi_\beta)},$$

which yields from Lemma 17

$$\begin{aligned} W_2\left(\mathcal{L}(\theta^\lambda_t), \pi_\beta\right) &\le \left(\sqrt{e^{3a}(\bar{C}_1 + \bar{C}_2 + \bar{C}_3)} + z_2\right)\lambda^{\frac{1}{4}} \\ &+ \sqrt{2\hat{c}}e^{-C_0\lambda t/2}\left(1 + \mathbb{E}\left[V_2\left(\theta_0\right)\right] + \int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(d\theta)\right)^{1/2} \\ &\le C_3\lambda^{\frac{1}{4}} + C_4 e^{-C_5 n} \end{aligned}$$

where

$$\begin{aligned} C_3 &:= \sqrt{e^{3a}(\bar{C}_1 + \bar{C}_2 + \bar{C}_3)} + z_2, \\ C_4 &:= \sqrt{2\hat{c}}\left(1 + \mathbb{E}\left[V_2\left(\theta_0\right)\right] + \int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(d\theta)\right)^{1/2}, \\ C_5 &= \frac{C_0}{2}. \end{aligned}$$

Therefore, we have for $m \in \mathbb{N}_0$, $0 < \lambda < \lambda_{\max}$,

$$W_2\left(\mathcal{L}(\theta^\lambda_{mT}), \pi_\beta\right) \le C_3\lambda^{\frac{1}{4}} + C_4 e^{-C_5 m},$$

and by setting $n = mT$ and using $-\lambda m \le -\lambda n/T \le -n\lambda$, we complete the proof.

□

Proof of Theorem 9. We begin by decomposing expected excess risk (11) as follows:

$$\mathbb{E}[u(\theta^\lambda_n)] - u(\theta^*) \le \mathbb{E}[u(\theta^\lambda_n)] - \mathbb{E}[u(Z_\infty)] + \mathbb{E}[u(Z_\infty)] - u(\theta^*) \tag{22}$$

where Z_∞ follows the target invariant measure π_β, i.e., $\mathcal{L}(Z_\infty) = \pi_\beta$. Let us focus on estimating the first part, $\mathbb{E}[u(\theta^\lambda_n)] - \mathbb{E}[u(Z_\infty)]$. Due to Remark 2, it follows that

$$\begin{aligned} |\nabla u(\theta)| &= |h(\theta)| \le (2^{2r}\mathbb{E}[K(X_0)] + \eta)(1 + |\theta|^{2r+1}) \\ &\le r_1(1 + |\theta|^{2r+1}), \end{aligned}$$

where $r_1 := 2^{2r}\mathbb{E}[K(X_0)] + \eta$. Then, one calculates that, for all $\theta, \theta \in \mathbb{R}^d$,

$$\begin{aligned} u(\theta) - u(\theta') &= \int_0^1 \langle \nabla u(t\theta + (1-t)\theta'),\, \theta - \theta' \rangle \mathrm{d}t \\ &\leq \int_0^1 \left(r_1 + r_1 2^{2r}(t^{2r+1}|\theta|^{2r+1} + (1-t)^{2r+1}|\theta'|^{2r+1}) \right) \mathrm{d}t |\theta - \theta'| \\ &\leq \left(r_1 + \frac{r_1 2^{2r}}{2r+2}|\theta|^{2r+1} + \frac{r_1 2^{2r}}{2r+2}|\theta'|^{2r+1} \right) |\theta - \theta'|. \end{aligned} \tag{23}$$

Let $\mathbf{P}$ denote the coupling between μ and ν that achieves $W_2(\mu, \nu)$ with $\mu = \mathcal{L}(\theta_n^\lambda)$ and $\nu = \mathcal{L}(Z_\infty)$, i.e., $W_2^2\left(\mathcal{L}(\theta_n^\lambda), \mathcal{L}(Z_\infty)\right) = \mathbb{E}_{\mathbf{P}}\left[|\theta_n^\lambda - Z_\infty|^2\right]$. By using (23) and Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} \mathbb{E}[u(\theta_n^\lambda)] - \mathbb{E}[u(Z_\infty)] &= \mathbb{E}_{\mathbf{P}}[u(\theta_n^\lambda) - u(Z_\infty)] \\ &\leq \mathbb{E}_{\mathbf{P}}\left[\left(r_1 + \frac{r_1 2^{2r}}{2r+2}|\theta_n^\lambda|^{2r+1} + \frac{r_1 2^{2r}}{2r+2}|Z_\infty|^{2r+1} \right) |\theta_n^\lambda - Z_\infty| \right] \\ &\leq \left(r_1 + \frac{r_1 2^{2r}}{2r+2}\sqrt{\mathbb{E}|\theta_n^\lambda|^{4r+2}} + \frac{r_1 2^{2r}}{2r+2}\sqrt{\mathbb{E}|Z_\infty|^{4r+2}} \right) W_2\Big(\mathcal{L}(\theta_n^\lambda), \pi_\beta \Big) \\ &\leq \left(r_1 + \frac{r_1 2^{2r}}{2r+2}\sqrt{\mathbb{E}|\theta_0^\lambda|^{4r+2} + \frac{A_{2r+1}}{\eta^2}} + \frac{r_1 2^{2r}}{2r+2}\sqrt{\mathbb{E}|Z_\infty|^{4r+2}} \right) \\ &\quad \times W_2\Big(\mathcal{L}(\theta_n^\lambda), \pi_\beta \Big) \\ &\leq C_6 W_2\Big(\mathcal{L}(\theta_n^\lambda), \pi_\beta \Big) \end{aligned} \tag{24}$$

where we have used Lemma 13 for the last inequality and the constant C_6 is given by

$$C_6 := r_1 + \frac{r_1 2^{2r}}{2r+2}\sqrt{\mathbb{E}|\theta_0^\lambda|^{4r+2} + \frac{A_{2r+1}}{\eta^2}} + \frac{r_1 2^{2r}}{2r+2}\sqrt{\mathbb{E}|Z_\infty|^{4r+2}}.$$

We take a similar approach in Raginsky et al. (2017) to estimate the second term in the RHS of (22). From Equation (3.18), (3.20) in Raginsky et al. (2017), we obtain

$$\begin{aligned} \mathbb{E}u(Z_\infty) - u(\theta^*) &\leq \frac{1}{\beta}\left(-\int_{\mathbb{R}^d} \frac{e^{-\beta u(\theta)}}{\Lambda} \log \frac{e^{-\beta u(\theta)}}{\Lambda} d\theta - \log \Lambda \right) - u^* \\ &\leq \frac{d}{2\beta} \log\left(\frac{2\pi e(B + d/\beta)}{Ad} \right) - \frac{\log \Lambda}{\beta} - u^* \end{aligned} \tag{25}$$

where $\Lambda = \int_{\mathbb{R}^d} e^{-\beta u(\theta)} d\theta$ is the normalizing constant.

Using (6), we obtain

$$\langle \theta^*,\, h(\theta^*) \rangle \geq A|\theta^*|^2 - B$$

which yields

$$|\theta^*|^2 \leq \sqrt{\frac{B}{A}}.$$

Moreover, for $w \in \mathbb{R}^d$, we have

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

From Remark 6, we further obtain

$$\begin{aligned} -\beta(u(\theta^*) - u(w)) &= \beta |u(\theta^*) - u(w)| \\ &\leq \beta \int_0^1 \frac{1}{t-1} |\langle h(w + t(\theta^* - w)) - h(\theta^*),\, w - \theta^* + t(\theta^* - w) \rangle| dt \\ &\leq \beta L_h \int_0^1 (1 + |w + t(\theta^* - w)| + |\theta^*|)^{2r+1} (1-t) |w - \theta^*|^2 dt \\ &\leq \beta L_h \int_0^1 (1 + |w| + |\theta^* - w| + |\theta^*|)^{2r+1} (1-t) |w - \theta^*|^2 dt \\ &= \beta L_h (1 + 2|\theta^*| + 2|\theta^* - w|)^{2r+1} \frac{|w - \theta^*|^2}{2} \end{aligned}$$

where we have used the elementary inequality $0 \leq |w| - |\theta^*| \leq |\theta^* - w|$ for the last inequality.

Define $R_0 := \max\{\sqrt{B/A}, \sqrt{2d/(\beta L_h)}\}$ and $\overline{\mathbf{B}}_r(p) = \{x \in \mathbb{R}^d | |x - p| > r\}$. Then, from the above inequality, one further calculates

$$\begin{aligned} \frac{\log \Lambda}{\beta} &= -u(\theta^*) + \frac{1}{\beta} \log \int_{\mathbb{R}^d} e^{\beta(u(\theta^*) - u(w))} dw \\ &\geq -u(\theta^*) + \frac{1}{\beta} \log \int_{\mathbb{R}^d} e^{-\beta L_h (1 + 2|\theta^*| + 2|\theta^* - w|)^{2r+1} \frac{|w - \theta^*|^2}{2}} dw \\ &\geq -u(\theta^*) + \frac{1}{\beta} \log \int_{\overline{\mathbf{B}}_{R_0}(\theta^*)} e^{-\beta L_h (1 + 4R_0)^{2r+1} \frac{|w - \theta^*|^2}{2}} dw \\ &= -u(\theta^*) + \frac{1}{\beta} \log \left[\left(\frac{2\pi}{\beta K} \right)^{d/2} \int_{\overline{\mathbf{B}}_{R_0}(\theta^*)} f_X(w) dw \right] \\ &\geq -u(\theta^*) + \frac{1}{\beta} \log \left(\frac{1}{2} \left(\frac{2\pi}{K\beta} \right)^{d/2} \right) \end{aligned} \tag{26}$$

where $K = L_h(1 + 4R_0)^{2r+1}$ and f_X is the density function of a multivariate normal variable X with mean θ^* and covariance $\frac{1}{K\beta} I_d$. Here, the last inequality is obtained from the

following inequality:

$$\begin{aligned}\int_{\overline{\mathbf{B}}_{R_0}(\theta^*)} f_X(w)dw &= P(|X-\theta^*|>R_0)\\ &= P\bigg(|X-\theta^*|>\sqrt{\frac{K\beta R_0^2}{d}}\sqrt{\frac{d}{K\beta}}\bigg)\\ &\leq \frac{d}{K\beta R_0^2}\\ &\leq \frac{1}{2(1+4R_0)^{2r+1}}\\ &\leq \frac{1}{2}.\end{aligned}$$

Combining (25) and (26), we derive

$$\begin{aligned}\mathbb{E}u(Z_\infty)-u(\theta^*) &\leq \frac{d}{2\beta}\log\bigg(\frac{2\pi e(B+d/\beta)}{Ad}\bigg)-\frac{1}{\beta}\log\bigg(\frac{1}{2}\bigg(\frac{2\pi}{K\beta}\bigg)^{d/2}\bigg)\\ &\leq \frac{1}{\beta}\bigg[\frac{d}{2}\log\bigg(\frac{Ke}{A}\bigg(\frac{B}{d}\beta+1\bigg)\bigg)+\log 2\bigg]. \qquad (27)\end{aligned}$$

Consequently, from (24) and (27), we derive

$$\begin{aligned}\mathbb{E}u(\theta_n^\lambda)-u(\theta^*) &\leq C_6 W_2(\mathcal{L}(\theta_n^\lambda,\pi_\beta))\\ &+ \frac{d}{2}\log\bigg(\frac{Ke}{A}\bigg(\frac{B}{d}\beta+1\bigg)\bigg)+\log 2.\end{aligned}$$

□

5. Conclusion and Discussion

This paper begins with an example which illustrates that local Lipschitz continuous gradients can cause serious convergence issues for popular adaptive optimization methods. Such issues manifest themselves as vanishing/exploding gradient phenomena. It proceeds by proposing a novel optimization framework, which is suitable for the fine tuning of neural network models by combining elements of the theory of Langevin SDEs, tamed algorithms and carefully designed boosting functions that handle sparse and super-linearly growing gradients. Further, a detailed convergence analysis of the newly proposed algorithm TheoPouLa is provided along with full theoretical guarantees for obtaining the best known convergence rates. Our experiments confirm that TheoPouLa outperforms other popular stochastic optimization methods.

We believe that this work opens a new door for stochastic (adaptive) optimization methods beyond the popular ADAM-type framework. Also, there is much room for improvement of our novel framework. For example, the improved performance can be further achieved by identifying more efficient taming and boosting functions, which demonstrates the potential of our framework.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801215 and the University of Edinburgh Data-Driven Innovation programme, part of the Edinburgh and South East Scotland City Region Deal.

A. Auxiliary Results

This section introduces some auxiliary results and their proofs, which are useful to obtain the main results.

Proof of Remark 2. By Assumption 2, it follows that for all $\theta \in \mathbb{R}^d$, $x \in \mathbb{R}^m$,

$$\begin{aligned} |G(\theta,x)| &\leq L_G(1+|x|)^\rho(1+|\theta|)^{q-1}|\theta| + |G(0,x)| \\ &\leq L_G(1+|x|)^\rho(1+|\theta|)^q + |G(0,x)|(1+|\theta|)^q \\ &\leq K_G(x)(1+|\theta|)^q, \end{aligned}$$

where $K_G(x) = L_G(1+|x|)^\rho + |G(0,x)|$. □

Proof of Remark 3. From the definition of H in (5), one obtains, for all $\theta \in \mathbb{R}^d$,

$$\begin{aligned} \langle \theta,\, h(\theta)\rangle &= \langle \theta,\, \mathbb{E}[G(\theta,X_0)]\rangle + \langle \theta,\, \eta\theta|\theta|^{2r}\rangle \\ &\geq \eta|\theta|^{2r+2} - 2^q\mathbb{E}[K_G(X_0)](1+|\theta|^{q+1}) \\ &\geq \eta|\theta|^{2r+2} - 2^q\mathbb{E}[1+K_G(X_0)](1+|\theta|^{q+1}). \end{aligned} \tag{28}$$

To prove (6), we want to show

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

for $A = 2^q\mathbb{E}[1+K_G(X_0)]$, $B = 3(2^{q+1}\mathbb{E}[1+K_G(X_0)])^{q+2}/\eta^{q+1}$.

One first observes that, for $|\theta| \geq 2A/\eta > 1$ with $\eta \in (0,1)$ and $r \geq q/2 \geq 1/2$,

$$\begin{aligned} \eta|\theta|^{2r+2} + B &\geq \frac{\eta}{2}|\theta|^3 + \frac{\eta}{2}|\theta|^{q+2} + \frac{3(2^{q+1}\mathbb{E}[1+K_G(X_0)])^{q+2}}{\eta^{q+1}} \\ &\geq 2^q\mathbb{E}[1+K_G(X_0)]|\theta|^2 + 2^q\mathbb{E}[1+K_G(X_0)]|\theta|^{q+1} + 2^q\mathbb{E}[1+K_G(X_0)] \\ &\geq A|\theta|^2 + 2^q\mathbb{E}[1+K_G(X_0)](1+|\theta|^{q+1}) \end{aligned} \tag{29}$$

Similarly, one can check that, for $|\theta| < 2A/\eta$,

$$\begin{aligned} \eta|\theta|^{2r+2} + B &\geq \frac{3(2^{q+1}\mathbb{E}[1+K_G(X_0)])^{q+2}}{\eta^{q+1}} \\ &\geq \eta\left(\frac{2^{q+1}\mathbb{E}[1+K_G(X_0)]}{\eta}\right)^3 + \eta\left(\frac{2^{q+1}\mathbb{E}[1+K_G(X_0)]}{\eta}\right)^{q+2} \\ &+ 2^q\mathbb{E}[1+K_G(X_0)] \\ &\geq 2^{q+1}\mathbb{E}[1+K_G(X_0)]|\theta|^2 + 2^{q+1}\mathbb{E}[1+K_G(X_0)]|\theta|^{q+1} + 2^q\mathbb{E}[1+K_G(X_0)] \\ &\geq A|\theta|^2 + 2^q\mathbb{E}[1+K_G(X_0)](1+|\theta|^{q+1}) \end{aligned} \tag{30}$$

Therefore, using (28), (29), and (30), we have for all $\theta \in \mathbb{R}^d$

$$\begin{aligned} \langle \theta,\, h(\theta)\rangle &\geq \eta|\theta|^{2r+2} - 2^q\mathbb{E}[1+K_G(X_0)](1+|\theta|^{q+1}) \\ &\geq A|\theta|^2 - B \end{aligned}$$

where $A = 2^q\mathbb{E}[1+K_G(X_0)]$, $B = 3(2^{q+1}\mathbb{E}[1+K_G(X_0)])^{q+2}/\eta^{q+1}$. □

Remark 21. *Let Assumption 2 holds and recall that the definitions of H and $H_{\lambda,c}$ are given in (5) and (9), respectively. Then, one obtains that, for all $\theta \in \mathbb{R}^d$, $x \in \mathbb{R}^m$, $i = 1, \ldots, d$,*

$$|H^{(i)}(\theta, x) - H^{(i)}_{\lambda,c}(\theta, x)| \leq \left(|G^{(i)}(\theta, x)|^2 + 1 + \eta|\theta^{(i)}||\theta|^{2r}\right)\sqrt{\lambda}. \tag{31}$$

Moreover, it follows that, for all $\theta \in \mathbb{R}^d$, $x \in \mathbb{R}^m$,

$$|H(\theta, x) - H_{\lambda,c}(\theta, x)|^2 \leq 9\left[8|K_G(x)|^4(1 + |\theta|^{4q}) + d + \eta^2|\theta|^{4r+2}\right]\lambda. \tag{32}$$

Proof of Remark 21. Recall the expressions of H and $H_{\lambda,c}$ in (5) and (9), respectively. The difference between the $H^{(i)}$ and $H^{(i)}_{\lambda,c}$ can be estimated by, for $i = 1, \ldots, d$,

$$\begin{aligned}
|H^{(i)}(\theta, x) - H^{(i)}_{\lambda,c}(\theta, x)| &\leq \left|G^{(i)}(\theta, x) - \frac{G^{(i)}(\theta, x)}{1 + \sqrt{\lambda}|G^{(i)}(\theta, x)|}\left(1 + \frac{\sqrt{\lambda}}{\varepsilon + |G^{(i)}(\theta, x)|}\right)\right| \\
&+ \left|\eta\theta^{(i)}|\theta|^{2r} - \eta\frac{\theta^{(i)}|\theta|^{2r}}{1 + \sqrt{\lambda}|\theta|^{2r}}\right| \\
&\leq |G^{(i)}(\theta, x)|\frac{\sqrt{\lambda}|G^{(i)}(\theta, x)|}{1 + \sqrt{\lambda}|G^{(i)}(\theta, x)|} \\
&+ \frac{\sqrt{\lambda}|G^{(i)}(\theta, x)|}{(1 + \sqrt{\lambda}|G^{(i)}(\theta, x)|)(\varepsilon + |G^{(i)}(\theta, x)|)} \\
&+ \eta|\theta^{(i)}||\theta|^{2r}\left|\frac{\sqrt{\lambda}|\theta|^{2r}}{1 + \sqrt{\lambda}|\theta|^{2r}}\right| \\
&\leq \sqrt{\lambda}|G^{(i)}(\theta, x)|^2 + \sqrt{\lambda} + \sqrt{\lambda}\eta|\theta^{(i)}||\theta|^{2r}.
\end{aligned}$$

By Remark 2 and Cauchy-Schwarz inequality, the above estimate further yields that

$$\begin{aligned}
|H(\theta, x) - H_{\lambda,c}(\theta, x)|^2 &= \sum_{i=1}^{d}\left(\sqrt{\lambda}|G^{(i)}(\theta, x)|^2 + \sqrt{\lambda} + \sqrt{\lambda}\eta|\theta^{(i)}||\theta|^{2r}\right)^2 \\
&\leq 9\lambda\sum_{i=1}^{d}\left[|G^{(i)}(\theta, x)|^4 + 1 + \eta^2|\theta^{(i)}|^2|\theta|^{4r}\right] \\
&\leq 9\lambda\left[\left(\sum_{i=1}^{d}|G^{(i)}(\theta, x)|^2\right)^2 + d + \eta^2|\theta|^{8r+2}\right] \\
&\leq 9\lambda\left[|G(\theta, x)|^4 + d + \eta^2|\theta|^{4r+2}\right] \\
&\leq 9\lambda\left[8|K_G(x)|^4(1 + |\theta|^{4q}) + d + \eta^2|\theta|^{4r+2}\right].
\end{aligned}$$

□

Remark 22. *Let Assumption 2 holds and recall that the definitions of H and $H_{\lambda,c}$ are given in* (5) *and* (9)*, respectively. Then, the growth of H can be estimated as follows: for all $\theta \in \mathbb{R}^d$, $x \in \mathbb{R}^m$,*

$$|H(\theta,x)|^2 \leq 4|K_G(x)|^2(1+|\theta|^{2q}) + 2\eta^2|\theta|^{4r+2}.$$

Moreover, one obtains, for all $\theta \in \mathbb{R}^d$, $x \in \mathbb{R}^m$,

$$|H_{\lambda,c}(\theta,x)|^2 \leq 18|K_G(x)|^2(1+|\theta|^{2q}) + 9\lambda d + 9\eta^2|\theta|^{4r+2}.$$

Proof of Remark 22. By Remark 2, one calculates that, for all $\theta \in \mathbb{R}^d$ and $x \in \mathbb{R}^m$,

$$\begin{aligned} |H(\theta,x)|^2 = |G(\theta,x) + \eta\theta|\theta|^{2r}|^2 &\leq 2|G(\theta,x)|^2 + 2\eta^2|\theta|^{4r+2} \\ &\leq 4|K_G(x)|^2(1+|\theta|^{2q}) + 2\eta^2|\theta|^{4r+2}, \end{aligned}$$

and

$$\begin{aligned} |H_{\lambda,c}(\theta,x)|^2 &= \sum_{i=1}^d \left(\frac{G^{(i)}(\theta,x)}{1+\sqrt{\lambda}|G^{(i)}(\theta,x)|}\left(1+\frac{\sqrt{\lambda}}{\varepsilon+|G^{(i)}(\theta,x)|}\right) + \eta\frac{\theta^{(i)}|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}}\right)^2 \\ &\leq \sum_{i=1}^d \left(\frac{|G^{(i)}(\theta,x)|}{1+\sqrt{\lambda}|G^{(i)}(\theta,x)|} + \frac{\sqrt{\lambda}|G^{(i)}(\theta,x)|}{(1+\sqrt{\lambda}|G^{(i)}(\theta,x)|)(\varepsilon+|G^{(i)}(\theta,x)|)} \right. \\ &+ \left. \eta\frac{|\theta^{(i)}|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}}\right)^2 \\ &\leq \sum_{i=1}^d \left(|G^{(i)}(\theta,x)| + \sqrt{\lambda} + \eta|\theta^{(i)}||\theta|^{2r}\right)^2 \\ &\leq 9\sum_{i=1}^d \left(|G^{(i)}(\theta,x)|^2 + \lambda + \eta^2|\theta^{(i)}|^2|\theta|^{4r}\right) \\ &\leq 9\left(|G(\theta,x)|^2 + \lambda d + \eta^2|\theta|^{4r+2}\right) \\ &\leq 18|K_G(x)|^2(1+|\theta|^{2q}) + 9\lambda d + 9\eta^2|\theta|^{4r+2}. \end{aligned}$$

□

B. Proofs of Lemmas in Appendix 4

Proof of Lemma 12. For each $i = 1, \ldots, d$ and fixed $\varepsilon > 0$, define

$$\widehat{G}^{(i)}_{\lambda,c}(\theta,x) = \frac{G^{(i)}(\theta,x)}{1+\sqrt{\lambda}|G^{(i)}(\theta,x)|}\left(1+\frac{\sqrt{\lambda}}{\varepsilon+|G^{(i)}(\theta,x)|}\right),$$

for all $\theta \in \mathbb{R}^d$, $x \in \mathbb{R}^m$, $0 < \lambda \leq \lambda_{\max}$. Denote by $\theta^{(i)}$ the i-th component of $\theta \in \mathbb{R}^d$ for $i = 1, \ldots, d$. One then observes that, for all $\theta \in \mathbb{R}^d$, $x \in \mathbb{R}^m$, $i = 1, \ldots, d$,

$$\begin{aligned} |\widehat{G}^{(i)}_{\lambda,c}(\theta, x)| &= \frac{|G^{(i)}(\theta, x)|}{1 + \sqrt{\lambda}|G^{(i)}(\theta, x)|} + \sqrt{\lambda}\frac{|G^{(i)}(\theta, x)|}{(1 + \sqrt{\lambda}|G^{(i)}(\theta, x)|)(\varepsilon + |G^{(i)}(\theta, x)|)} \\ &\leq \frac{1}{\sqrt{\lambda}} + \sqrt{\lambda}\frac{|G^{(i)}(\theta, x)|/\varepsilon}{1 + |G^{(i)}(\theta, x)|/\varepsilon} \\ &\leq \frac{1}{\sqrt{\lambda}} + \sqrt{\lambda}. \end{aligned} \tag{33}$$

By using Cauchy-Schwartz inequality and (33), one can further calculate that for all $\theta \in \mathbb{R}^d$, $x \in \mathbb{R}^m$,

$$\begin{aligned} \langle \theta,\, H_{\lambda,c}(\theta, x)\rangle &= \sum_{i=1}^{d} \theta^{(i)} \cdot \widehat{G}^{(i)}_{\lambda,c}(\theta, x) + \eta\frac{|\theta|^{2r+2}}{1 + \sqrt{\lambda}|\theta|^{2r}} \\ &\geq \sum_{i=1}^{d} |\theta^{(i)}|\Big(- \frac{1}{\sqrt{\lambda}} - \sqrt{\lambda}\Big) + \eta\frac{|\theta|^{2r+2}}{1 + \sqrt{\lambda}|\theta|^{2r}} \\ &\geq -\Big(\frac{1}{\sqrt{\lambda}} + \sqrt{\lambda}\Big)\sqrt{d}|\theta| + \eta\frac{|\theta|^{2r+2}}{1 + \sqrt{\lambda}|\theta|^{2r}}, \end{aligned} \tag{34}$$

which implies that

$$-\frac{2\lambda}{|\theta^{\lambda}_n|^2}\mathbb{E}\left[\langle \theta^{\lambda}_n,\, H_{\lambda,c}(\theta^{\lambda}_n, X_{n+1})\rangle \middle| \theta^{\lambda}_n\right] \leq \frac{2\sqrt{d}}{|\theta^{\lambda}_n|}\Big(\sqrt{\lambda} + \lambda^{\frac{3}{2}}\Big) - \frac{2\eta\lambda|\theta^{\lambda}_n|^{2r}}{1 + \sqrt{\lambda}|\theta^{\lambda}_n|^{2r}}. \tag{35}$$

Moreover, using (33), it is shown that, for all $\theta \in \mathbb{R}^d, x \in \mathbb{R}^m$,

$$\begin{aligned} |H_{\lambda,c}(\theta, x)|^2 &= \sum_{i=1}^{d}\left(\widehat{G}^{(i)}_{\lambda,c}(\theta, x) + \eta\frac{\theta^{(i)}|\theta|^{2r}}{1 + \sqrt{\lambda}|\theta|^{2r}}\right)^2 \\ &\leq \sum_{i=1}^{d}\left(2|\widehat{G}^{(i)}_{\lambda,c}(\theta, x)|^2 + 2\eta^2\frac{|\theta^{(i)}|^2|\theta|^{4r}}{(1 + \sqrt{\lambda}|\theta|^{2r})^2}\right) \\ &\leq 4d\Big(\frac{1}{\lambda} + \lambda\Big) + 2\eta^2|\theta|^2\frac{|\theta|^{4r}}{(1 + \sqrt{\lambda}|\theta|^{2r})^2}, \end{aligned} \tag{36}$$

which yields that

$$\frac{\lambda^2}{|\theta^{\lambda}_n|^2}\mathbb{E}\left[|H_{\lambda,c}(\theta^{\lambda}_n, X_{n+1})|^2 \middle| \theta^{\lambda}_n\right] \leq 4\lambda d\frac{(1 + \lambda^2)}{|\theta^{\lambda}_n|^2} + 2\lambda\eta^2. \tag{37}$$

Combining (35) and (37), one calculates that

$$\begin{aligned}
\mathbb{E}\left[|\theta_{n+1}^{\lambda}|^2|\theta_n^{\lambda}\right] &= \mathbb{E}\left[|\theta_n^{\lambda} - \lambda H_{\lambda,c}(\theta_n^{\lambda}, X_{n+1}) + \sqrt{2\lambda\beta^{-1}}\xi_{n+1}|^2|\theta_n^{\lambda}\right] \\
&= |\theta_n^{\lambda}|^2\bigg(1 - \frac{2\lambda}{|\theta_n^{\lambda}|^2}\mathbb{E}\left[\langle\theta_n^{\lambda},\, H_{\lambda,c}(\theta_n^{\lambda}, X_{n+1})\rangle\bigg|\theta_n^{\lambda}\right] \\
&+ \frac{\lambda^2}{|\theta_n^{\lambda}|^2}\mathbb{E}\left[|H_{\lambda,c}(\theta_n^{\lambda}, X_{n+1})|^2\bigg|\theta_n^{\lambda}\right]\bigg) + \frac{2\lambda d}{\beta} \\
&\leq |\theta_n^{\lambda}|^2\bigg[1 + \frac{2\sqrt{d}}{|\theta_n^{\lambda}|}\left(\sqrt{\lambda} + \lambda^{\frac{3}{2}}\right) - \frac{2\eta\lambda|\theta_n^{\lambda}|^{2r}}{1+\sqrt{\lambda}|\theta_n^{\lambda}|^{2r}} \\
&+ 4\lambda d\frac{(1+\lambda^2)}{|\theta_n^{\lambda}|^2} + 2\lambda\eta^2\bigg] + \frac{2\lambda d}{\beta} \\
&= |\theta_n^{\lambda}|^2\left(1 - f^{\lambda}(\theta_n^{\lambda})\right) + \frac{2\lambda d}{\beta}, \qquad (38)
\end{aligned}$$

where $f^{\lambda}(\theta) := -\frac{2\sqrt{d}}{|\theta|}\left(\sqrt{\lambda} + \lambda^{\frac{3}{2}}\right) + \frac{2\eta\lambda|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}} - 4\lambda d\frac{(1+\lambda^2)}{|\theta|^2} - 2\lambda\eta^2$ for all $\theta \in \mathbb{R}^d \setminus \mathbf{0}$. We note that, for all $0 < \lambda \leq \lambda_{\max} \leq (4\eta^2)^{-1}$,

$$\lim_{|\theta|\to\infty} f^{\lambda}(\theta) = 2\eta\sqrt{\lambda}(1 - \sqrt{\lambda}\eta) \geq \eta\sqrt{\lambda} > 0.$$

In addition, using the fact that $f(s) := s/(1 + \sqrt{\lambda}s)$ is non-decreasing for all $s \geq 0$, one can choose $M_0 > 0$ such that

$$f^{\lambda}(\theta) \geq \eta\sqrt{\lambda}(1 - \sqrt{\lambda}\eta) \geq \frac{\eta\sqrt{\lambda}}{2}, \qquad (39)$$

for all $\theta \geq M_0$, $0 < \lambda \geq \lambda_{\max}$. Therefore, from (38) and (39), we have that

$$\mathbb{E}\left[|\theta_{n+1}^{\lambda}|^2\mathbf{1}_{|\theta_n^{\lambda}|\geq M_0}\bigg|\theta_n^{\lambda}\right] \leq |\theta_n^{\lambda}|^2\left(1 - \frac{\eta\sqrt{\lambda}}{2}\right) + \frac{2\lambda d}{\beta}. \qquad (40)$$

Let us consider the case of $|\theta_n^{\lambda}| < M_0$. From (34) and (36), we have for all $|\theta| < M_0$, $x \in \mathbb{R}^m$,

$$\begin{aligned}
-2\lambda\langle\theta,\, H_{\lambda,c}(\theta, x)\rangle + \lambda^2|H_{\lambda,c}(\theta, x)|^2|\theta] &\leq 2\left(\sqrt{\lambda} + \lambda^{\frac{3}{2}}\right)\sqrt{d}M_0 \\
&+ 2\eta\sqrt{\lambda}M_0^2 + 4d(\lambda + \lambda^3) + 2\eta^2\lambda M_0^2.
\end{aligned}$$

The above estimate directly yields that, for $0 < \lambda \geq \lambda_{\max}$,

$$\begin{aligned}
\mathbb{E}\left[|\theta_{n+1}^{\lambda}|^2\mathbf{1}_{|\theta_n^{\lambda}|<M_0}\bigg|\theta_n^{\lambda}\right] &\leq |\theta_n^{\lambda}|^2 + \frac{2\lambda d}{\beta} + 2\left(\sqrt{\lambda} + \lambda^{\frac{3}{2}}\right)\sqrt{d}M_0 + 2\eta\sqrt{\lambda}M_0^2 \\
&+ 4d(\lambda + \lambda^3) + 2\eta^2\lambda M_0^2 \\
&\leq \left(1 - \frac{\eta\sqrt{\lambda}}{2}\right)|\theta_n^{\lambda}|^2 \\
&+ \sqrt{\lambda}\left(\frac{2d}{\beta} + 4\sqrt{d}M_0 + \frac{5\eta}{2}M_0^2 + 8d + 2\eta^2M_0^2\right). \qquad (41)
\end{aligned}$$

Consequently, (40) and (41) yield that

$$\mathbb{E}\left[|\theta^\lambda_{n+1}|^2\middle|\theta^\lambda_n\right] \leq \left(1-\frac{\eta\sqrt{\lambda}}{2}\right)|\theta^\lambda_n|^2+\sqrt{\lambda}\left(\frac{2d}{\beta}+4\sqrt{d}M_0+\frac{5\eta}{2}M_0^2+8d+2\eta^2M_0^2\right),$$

and

$$\begin{aligned}\mathbb{E}\left[|\theta^\lambda_{n+1}|^2\right] &\leq \left(1-\frac{\eta\sqrt{\lambda}}{2}\right)^n\mathbb{E}|\theta^\lambda_0|^2\\ &+ \sqrt{\lambda}\left(\frac{2d}{\beta}+4\sqrt{d}M_0+\frac{5\eta}{2}M_0^2+8d+2\eta^2M_0^2\right)\sum_{j=0}^{\infty}\left(1-\frac{\eta\sqrt{\lambda}}{2}\right)^j\\ &\leq \left(1-\frac{\eta\sqrt{\lambda}}{2}\right)^n\mathbb{E}|\theta^\lambda_0|^2+\left(\frac{4d}{\beta\eta}+\frac{8\sqrt{d}M_0}{\eta}+5M_0^2+\frac{16d}{\eta}+4\eta M_0^2\right).\end{aligned}$$

□

Proof of Lemma 13. For any integer $p\geq 2$, $n\in\mathbb{N}_0$, $|\theta^\lambda_{n+1}|^{2p}$ is written as

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

where $\Delta_n=\theta^\lambda_n-\lambda H_{\lambda,c}(\theta^\lambda_n,X_{n+1})$. Then, we obtain

$$\begin{aligned}\mathbb{E}[|\theta^\lambda_{n+1}|^{2p}|\theta^\lambda_n] &= \mathbb{E}\left[\left(|\Delta_n|^2+\left|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\right|^2+2\langle\Delta_n,\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\rangle\right)^p\middle|\theta^\lambda_n\right]\\ &= \mathbb{E}[|\Delta_n|^{2p}|\theta^\lambda_n]+2p\mathbb{E}\left[|\Delta_n|^{2p-2}\langle\Delta_n,\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\rangle\middle|\theta^\lambda_n\right]\\ &+ \sum_{\substack{k_1+k_2+k_3=p\\ \{k_1\neq p-1\}\cap\{k_2\neq 1\}\\ \{k_1\neq p\}}}\frac{p!}{k_1!k_2!k_3!}\mathbb{E}\left[|\Delta_n|^{2k_1}\left|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\right|^{k_2}\right.\\ &\times \left.\left|2\langle\Delta_n,\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\rangle\right|^{k_3}\middle|\theta^\lambda_n\right]\\ &\leq \mathbb{E}[|\Delta_n|^{2p}|\theta^\lambda_n]+\sum_{k=2}^{2p}\binom{2p}{k}\mathbb{E}\left[|\Delta_n|^{2p-k}\left|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\right|^k\middle|\theta^\lambda_n\right] \qquad (42)\end{aligned}$$

where the above inequality follows from the result in (Chau et al., 2019, Lemma A.3). Using the fact that ξ_{n+1} is independent of Δ_n and θ_n^λ, one further calculates that

$$\begin{aligned}
\mathbb{E}[|\theta_{n+1}^\lambda|^{2p}|\theta_n^\lambda] &\leq \mathbb{E}[|\Delta_n|^{2p}|\theta_n^\lambda] + \sum_{l=0}^{2p-2}\binom{2p}{l+2}\mathbb{E}\left[|\Delta_n|^{2p-2-l}\left|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\right|^{l+2}\middle|\theta_n^\lambda\right] \\
&= \mathbb{E}[|\Delta_n|^{2p}|\theta_n^\lambda] \\
&+ \sum_{l=0}^{2p-2}\frac{2p(2p-1)}{(l+2)(l+1)}\binom{2p-2}{l}\mathbb{E}\left[|\Delta_n|^{2p-2-l}\left|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\right|^{l+2}\middle|\theta_n^\lambda\right] \\
&\leq \mathbb{E}[|\Delta_n|^{2p}|\theta_n^\lambda] \\
&+ p(2p-1)\mathbb{E}\left[\left(|\Delta_n| + \left|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\right|\right)^{2p-2}\left|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\right|^2\middle|\theta_n^\lambda\right] \\
&\leq \mathbb{E}[|\Delta_n|^{2p}|\theta_n^\lambda] \\
&+ 2^{2p-3}p(2p-1)\left(\mathbb{E}[|\Delta_n|^{2p-2}|\theta_n^\lambda]\frac{2\lambda d}{\beta} + \left(\frac{2\lambda}{\beta}\right)^p\mathbb{E}|\xi_{n+1}|^{2p}\right).
\end{aligned}$$

Define $|\Delta_n|^2 = |\theta_n^\lambda|^2 + r_n$ where $r_n = -2\lambda\langle\theta_n^\lambda, H_{\lambda,c}(\theta_n^\lambda, X_{n+1})\rangle + \lambda^2|H_{\lambda,c}(\theta_n^\lambda, X_{n+1})|^2$ to write

$$\begin{aligned}
\mathbb{E}\left[|\Delta_n|^{2p}|\theta_n^\lambda\right] &= \sum_{k=0}^{p}\binom{p}{k}|\theta_n^\lambda|^{2(p-k)}\mathbb{E}\left[r_n^k|\theta_n^\lambda\right] \\
&= |\theta_n^\lambda|^{2p} + p|\theta_n^\lambda|^{2p-2}\mathbb{E}\left[r_n|\theta_n^\lambda\right] + \sum_{k=2}^{p}\binom{p}{k}|\theta_n^\lambda|^{2(p-k)}\mathbb{E}\left[r_n^k|\theta_n^\lambda\right]. \qquad (43)
\end{aligned}$$

Now, we focus on the case where $|\theta_n^\lambda| > M$ where

$$M := \max\left\{M_0, 1, \frac{4d}{(2-\eta)\eta}, \frac{2\sqrt{d}}{\eta(2-\eta)}, \frac{2^{2p-2}p(2p-1)d}{\eta\beta}\right\}.$$

Recall that M_0 is defined in the proof of Lemma 12. We need the following relations to estimate the moments of r_n: for all $x \in \mathbb{R}^d$, $0 < \lambda \leq \lambda_{\max}$, $0 < \eta < 1$, $|\theta| \geq M$,

$$\begin{aligned}
\lambda^2|H_{\lambda,c}(\theta,x)|^2 &\leq 4d(\lambda+\lambda^3) + 2\eta^2\lambda|\theta|^2\frac{\lambda|\theta|^{4r}}{(1+\sqrt{\lambda}|\theta|^{2r})^2} \\
&\leq 4d\lambda(1+\lambda^2) + 2\eta^2\lambda|\theta|^2 \\
&\leq 4d\lambda(1+\lambda^2)|\theta| + 2\eta^2\lambda|\theta|^2 \\
&\leq 2\sqrt{\lambda}\eta\left(\frac{4d}{|\theta|\eta} + \eta\right)|\theta|^2 \\
&\leq 2\sqrt{\lambda}\eta\left(\frac{4d}{M\eta} + \eta\right)|\theta|^2 \\
&\leq 4\sqrt{\lambda}\eta|\theta|^2, \qquad (44)
\end{aligned}$$

where we have used the inequality (36) and

$$M > \frac{4d}{(2-\eta)\eta} \Leftrightarrow \left(\frac{4d}{M\eta} + \eta\right) < 2.$$

Notice that $\frac{4d}{(2-\eta)\eta}$ is finite due to $\lambda_{\max}$ being less than $\frac{1}{4\eta^2}$. Moreover, from (34), we have the following inequality, for $0 < \lambda \leq \lambda_{\max}$,

$$\begin{aligned}
|2\lambda\langle\theta,\, H_{\lambda,c}(\theta,x)\rangle| &\leq 2(\sqrt{\lambda}+\lambda^{1.5})\sqrt{d}|\theta| + 2\eta\sqrt{\lambda}|\theta|^2\frac{\sqrt{\lambda}|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}} \\
&\leq 2\sqrt{\lambda}(1+\lambda)\sqrt{d}|\theta| + 2\eta\sqrt{\lambda}|\theta|^2 \\
&\leq 2\sqrt{\lambda}\eta\left(\frac{2\sqrt{d}}{|\theta|\eta}+\eta\right)|\theta|^2 \\
&\leq 2\sqrt{\lambda}\eta\left(\frac{2\sqrt{d}}{M\eta}+\eta\right)|\theta|^2 \\
&\leq 4\sqrt{\lambda}\eta|\theta|^2, \qquad (45)
\end{aligned}$$

where the last inequality holds since

$$M > \frac{2\sqrt{d}}{\eta(2-\eta)} \Leftrightarrow \left(\frac{2\sqrt{d}}{M\eta}+\eta\right) \leq 2.$$

Thus, r_n^k can be written as

$$\begin{aligned}
\mathbb{E}[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}|r_n|^k|\theta_n^\lambda] &= \mathbb{E}\Bigg[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}\Big(-2\lambda\langle\theta_n^\lambda,\, H_{\lambda,c}(\theta_n^\lambda,X_{n+1})\rangle \\
&+ \lambda^2|H_{\lambda,c}(\theta_n^\lambda,X_{n+1})|^2\Big)^k\Big|\theta_n^\lambda\Bigg] \\
&\leq \mathbb{E}\Bigg[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}\Big(|2\lambda\langle\theta_n^\lambda,\, H_{\lambda,c}(\theta_n^\lambda,X_{n+1})\rangle| \\
&+ \lambda^2|H_{\lambda,c}(\theta_n^\lambda,X_{n+1})|^2\Big)^k\Big|\theta_n^\lambda\Bigg] \\
&\leq \mathbb{E}\Bigg[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}(8\sqrt{\lambda}\eta|\theta_n^\lambda|^2)^k\Big|\theta_n^\lambda\Bigg] \\
&\leq \lambda^{\frac{k}{2}}(8\eta)^k|\theta_n^\lambda|^{2k}.
\end{aligned}$$

Moreover, (39) implies that

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

or, equivalently,

$$p|\theta_n^\lambda|^{2p-2}\mathbb{E}[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}r_n|\theta_n^\lambda] \leq -p\frac{\eta\sqrt{\lambda}}{2}|\theta_n^\lambda|^{2p}. \qquad (46)$$

Using (46), the L_{2p}-norm of Δ_n conditional on $\theta_n^\lambda > M$ is given by

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}|\Delta_n|^{2p}\middle|\theta_n^\lambda\right] &\leq |\theta_n^\lambda|^{2p} + p|\theta_n^\lambda|^{2p-2}\mathbb{E}[\mathbf{1}_{\{\theta_n^\lambda>M\}}r_n|\theta_n^\lambda] \\
&+ \sum_{k=2}^{p}\binom{p}{k}|\theta_n^\lambda|^{2(p-k)}\mathbb{E}[\mathbf{1}_{\{\theta_n^\lambda>M\}}|r_n|^k|\theta_n^\lambda] \\
&\leq |\theta_n^\lambda|^{2p} - p\frac{\eta\sqrt{\lambda}}{2}|\theta_n^\lambda|^{2p} + \sum_{k=2}^{p}\binom{p}{k}|\theta_n^\lambda|^{2(p-k)}\lambda^{\frac{k}{2}}(8\eta)^k|\theta_n^\lambda|^{2k} \\
&\leq |\theta_n^\lambda|^{2p} - p\frac{\eta\sqrt{\lambda}}{2}|\theta_n^\lambda|^{2p} + |\theta_n^\lambda|^{2p}\sum_{k=2}^{p}\binom{p}{k}\lambda^{\frac{k}{2}}(8\eta)^k. \qquad (47)
\end{aligned}$$

Moreover, it follows that, for $0 < \lambda \leq \lambda_{\max}$,

$$\lambda \leq \frac{1}{(2^7\eta_p\mathcal{C}_{\lceil\frac{p}{2}\rceil})^2} = \frac{1}{2^8(8\eta)^2{}_p\mathcal{C}^2_{\lceil\frac{p}{2}\rceil}} \leq \frac{1}{2^{\frac{8}{k-1}}(8\eta)^2({}_p\mathcal{C}^2_{\lceil\frac{p}{2}\rceil})^{\frac{2}{k-1}}},$$

which is equivalent to

$$\begin{aligned}
\lambda^{\frac{k-1}{2}} &\leq \frac{1}{2^4(8\eta)^{k-1}{}_p\mathcal{C}_{\lceil\frac{p}{2}\rceil}} \\
&= \frac{\eta}{2(8\eta)^k{}_p\mathcal{C}_{\lceil\frac{p}{2}\rceil}},
\end{aligned}$$

for all $k \in [2,p]\cap\mathbb{N}$. Then, one observes the following inequality

$$\begin{aligned}
\sum_{k=2}^{p}{}_p\mathcal{C}_k\lambda^{\frac{k}{2}}(8\eta)^k &\leq \sum_{k=2}^{p}{}_p\mathcal{C}_{\lceil\frac{p}{2}\rceil}\lambda^{\frac{k}{2}}(8\eta)^k \\
&\leq \frac{1}{2}\sum_{k=2}^{p}\sqrt{\lambda}\eta \\
&= \frac{p-2}{2}\sqrt{\lambda}\eta,
\end{aligned}$$

to obtain

$$\mathbb{E}\left[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}|\Delta_n|^{2p}\middle|\theta_n^\lambda\right] \leq (1-\eta\sqrt{\lambda})|\theta_n^\lambda|^{2p}, \qquad (48)$$

and

$$\mathbb{E}\left[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}|\Delta_n|^{2p-2}\middle|\theta_n^\lambda\right] \leq (1-\eta\sqrt{\lambda})|\theta_n^\lambda|^{2(p-2)} \leq \frac{1}{M^2}(1-\eta\sqrt{\lambda})|\theta_n^\lambda|^{2p}. \qquad (49)$$

By combining (42), (49) and (48), we derive

$$\begin{aligned}\mathbb{E}[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}|\theta_{n+1}^\lambda|^{2p}|\theta_n^\lambda] &\leq (1-\eta\sqrt{\lambda})|\theta_n^\lambda|^{2p}\\ &+ \frac{2^{2p-2}p(2p-1)\lambda d}{M^2\beta}(1-\eta\sqrt{\lambda})|\theta_n^\lambda|^{2p}\\ &+ 2^{2p-3}p(2p-1)\left(\frac{2\lambda}{\beta}\right)^p\mathbb{E}|\xi_{n+1}|^{2p}\\ &\leq (1-\eta\sqrt{\lambda})\left(1+\frac{2^{2p-2}p(2p-1)\lambda d}{M^2\beta}\right)|\theta_n^\lambda|^{2p}\\ &+ 2^{2p-3}p(2p-1)\left(\frac{2\lambda}{\beta}\right)^p\mathbb{E}|\xi_{n+1}|^{2p}\\ &\leq (1-\eta^2\lambda)|\theta_n^\lambda|^{2p}+2^{2p-3}p(2p-1)\left(\frac{2\lambda}{\beta}\right)^p\mathbb{E}|\xi_{n+1}|^{2p}, \quad (50)\end{aligned}$$

where we used the fact that $M \geq \frac{2^{2p-2}p(2p-1)d}{\eta\beta}$ for the last inequality.

Consider the case of $|\theta_n^\lambda| \leq M$. By observing that from (36)

$$\mathbf{1}_{\{|\theta|\leq M\}}\lambda^2|H_{\lambda,c}(\theta,x)|^2 \leq \lambda\left(8d+2\eta^2M^2\right),$$

and

$$\begin{aligned}\mathbf{1}_{\{|\theta|\leq M\}}|2\lambda\langle\theta,\, H_{\lambda,c}(\theta,x)\rangle| &\leq 2\lambda\sqrt{|\theta|}\sqrt{|H_{\lambda,c}(\theta,x)|}\\ &\leq 2\lambda\sqrt{M}\sqrt{|G(\theta,x)|+d\sqrt{\lambda}+2\eta M^{2r+1}}\\ &\leq 2\lambda\sqrt{M}\sqrt{|K_G(x)|(1+M^q)+d\sqrt{\lambda}+2\eta M^{2r+1}},\end{aligned}$$

it can be shown that

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}_{\{|\theta_n^\lambda|\leq M\}}|r_n|^k\middle|\theta_n^\lambda\right] &= \mathbb{E}\left[\mathbf{1}_{\{|\theta_n^\lambda|\leq M\}}\left(|2\lambda\langle\theta_n^\lambda,\, H_{\lambda,c}(\theta_n^\lambda,X_{n+1})\rangle|\right.\right.\\ &+ \left.\left.\lambda^2|H_{\lambda,c}(\theta_n^\lambda,X_{n+1})|^2\right)^k\middle|\theta_n^\lambda\right]\\ &\leq \mathbb{E}\left[\mathbf{1}_{\{|\theta_n^\lambda|\leq M\}}\left(2\lambda\sqrt{M}\sqrt{K_G(X_{n+1})(1+M^q)+d+2\eta M^{2r+1}}\right.\right.\\ &+ \left.\left.\lambda\left(8d+2\eta^2M^2\right)\right)^k\middle|\theta_n^\lambda\right]\\ &\leq \widetilde{D}_k\lambda^k,\end{aligned}$$

where $\widetilde{D}_k = 2^{k-1}\left((2\sqrt{M})^k(\mathbb{E}[K_G(X_0)](1+M^q)+d+2\eta M^{2r+1})^{k/2}+(8d+2\eta^2M^2)^k\right)$.
Hence, one calculates that

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}_{\{|\theta_n^\lambda|\leq M\}}|\Delta_n|^{2p}\middle|\theta_n^\lambda\right] &\leq |\theta_n^\lambda|^{2p}+\sum_{k=1}^p\binom{p}{k}|\theta_n^\lambda|^{2(p-k)}\mathbb{E}[\mathbf{1}_{\{\theta_n^\lambda\leq M\}}|r_n|^k|\theta_n^\lambda]\\ &\leq (1-\eta^2\lambda)|\theta_n^\lambda|^{2p}+\eta^2\lambda M^{2p}+M^{2p}\lambda\sum_{k=1}^p\binom{p}{k}\lambda^{k-1}\widetilde{D}_k,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}_{\{\theta_n^\lambda \le M\}}|\Delta_n|^{2p-2}\middle|\theta_n^\lambda\right] &\le \sum_{k=0}^{p-1}\binom{p-1}{k}|\theta_n^\lambda|^{2(p-1-k)}\mathbb{E}[\mathbf{1}_{\{|\theta_n^\lambda|\le M\}}|r_n|^k|\theta_n^\lambda] \\ &\le M^{2p-2}\sum_{k=0}^{p-1}\binom{p}{k}\widetilde{D}_k\lambda^k.\end{aligned}$$

Consequently, we obtain

$$\begin{aligned}\mathbb{E}[\mathbf{1}_{\{|\theta_n^\lambda|\le M\}}|\theta_{n+1}^\lambda|^{2p}|\theta_n^\lambda] &\le (1-\eta^2\lambda)|\theta_n^\lambda|^{2p} + \eta^2\lambda M^{2p} + \lambda M^{2p}\sum_{k=1}^{p}\binom{p}{k}\lambda^{k-1}\widetilde{D}_k \\ &+ \frac{\lambda d}{\beta}2^{2p-2}p(2p-1)M^{2p-2}\sum_{k=0}^{p-1}\binom{p}{k}\lambda^k\widetilde{D}_k \qquad (51) \\ &+ 2^{2p-3}p(2p-1)\left(\frac{2\lambda}{\beta}\right)^p\mathbb{E}|\xi_{n+1}|^{2p}.\end{aligned}$$

By defining

$$\begin{aligned}A_p &= \eta^2M^{2p} + M^{2p}\sum_{k=1}^{p}\binom{p}{k}\widetilde{D}_k \\ &+ 2^{2p-3}p(2p-1)\left(\frac{2dM^{2p-2}}{\beta}\sum_{k=0}^{p-1}\binom{p}{k}\lambda^k\widetilde{D}_k + \frac{2}{\beta}\left(\frac{2}{\beta}\right)^{p-1}d^p(2p-1)!!\right),\end{aligned}$$

we conclude that

$$\begin{aligned}\mathbb{E}|\theta_{n+1}^\lambda|^{2p} &\le (1-\eta^2\lambda)\mathbb{E}|\theta_n^\lambda|^{2p} + \lambda A_p \\ &\le (1-\eta^2\lambda)^n\mathbb{E}|\theta_0^\lambda|^{2p} + \lambda A_p\sum_{j=0}^{\infty}(1-\eta^2\lambda)^j \\ &\le (1-\eta^2\lambda)^n\mathbb{E}|\theta_0^\lambda|^{2p} + \frac{A_p}{\eta^2}.\end{aligned}$$

□

Proof of Lemma 16. For $p \ge 1$, $0 < \lambda \le \lambda_{\max}$, $t \in (nT, (n+1)T]$, $n \in \mathbb{N}_0$, Itô's formula yields that

$$\begin{aligned}\mathbb{E}[V_p(\bar{\zeta}_t^{\lambda,n})] &= \mathbb{E}[V_p(\bar{\theta}_{nT}^\lambda)] + \int_{nT}^{t}\mathbb{E}\left[\lambda\frac{\Delta V_p(\bar{\zeta}_s^{\lambda,n})}{\beta} - \lambda\langle h(\bar{\zeta}_s^{\lambda,n}), \nabla V_p(\bar{\zeta}_s^{\lambda,n})\rangle\right]\mathrm{d}s \\ &\quad + \mathbb{E}\left[\int_{nT}^{t}\left\langle\nabla V_p(\bar{\zeta}_s^{\lambda,n}), \sqrt{2\lambda\beta^{-1}}\,\mathrm{d}B_s^\lambda\right\rangle\right] \\ &= \mathbb{E}[V_p(\bar{\theta}_{nT}^\lambda)] + \int_{nT}^{t}\mathbb{E}\left[\lambda\frac{\Delta V_p(\bar{\zeta}_s^{\lambda,n})}{\beta} - \lambda\langle h(\bar{\zeta}_s^{\lambda,n}), \nabla V_p(\bar{\zeta}_s^{\lambda,n})\rangle\right]\mathrm{d}s. \qquad (52)\end{aligned}$$

Then, differentiating both sides of (52) and applying Lemma 14, we have

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

Therefore, we have the following inequality:

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

By setting $p = 4$ and applying Lemma 15, we obtain the desired result:

$$\begin{aligned}\mathbb{E}[V_4(\bar{\zeta}_t^{\lambda,n})] &\leq e^{-\lambda(t-nT)\bar{c}(4)}\mathbb{E}[V_4(\bar{\theta}_{nT}^{\lambda})] + \frac{\tilde{c}(4)}{\bar{c}(4)}\left(1 - e^{-\lambda\bar{c}(4)(t-nT)}\right) \\ &\leq 2 + 2\mathbb{E}[|\theta_0|^4] + 2\frac{A_4}{\eta^2} + \frac{\tilde{c}(4)}{\bar{c}(4)}.\end{aligned}$$

□

Proof of Lemma 18. We begin by applying Itô's formula to observe, for all $n \in \mathbb{N}_0$ and $t \in (nT, (n+1)T]$,

$$\begin{aligned}W_2^2\left(\mathcal{L}(\bar{\theta}_t^{\lambda}), \mathcal{L}(\bar{\zeta}_t^{\lambda,n})\right) &\leq \mathbb{E}\left[|\bar{\theta}_t^{\lambda} - \bar{\zeta}_t^{\lambda,n}|^2\right] \\ &= -2\lambda\int_{nT}^{t}\mathbb{E}\left[\langle\bar{\zeta}_s^{\lambda,n} - \bar{\theta}_s^{\lambda}, h(\bar{\zeta}_s^{\lambda,n}) - H_\lambda(\bar{\theta}_{\lfloor s\rfloor}^{\lambda}, X_{\lceil s\rceil})\rangle\right]\mathrm{d}s \\ &= -2\lambda\int_{nT}^{t}\mathbb{E}\left[\langle\bar{\zeta}_s^{\lambda,n} - \bar{\theta}_s^{\lambda},\, h(\bar{\zeta}_s^{\lambda,n}) - h(\bar{\theta}_s^{\lambda})\rangle\right]\mathrm{d}s \\ &\quad - 2\lambda\int_{nT}^{t}\mathbb{E}\left[\langle\bar{\zeta}_s^{\lambda,n} - \bar{\theta}_s^{\lambda},\, h(\bar{\theta}_s^{\lambda}) - h(\bar{\theta}_{\lfloor s\rfloor}^{\lambda})\rangle\right]\mathrm{d}s \\ &\quad - 2\lambda\int_{nT}^{t}\mathbb{E}\left[\langle\bar{\zeta}_s^{\lambda,n} - \bar{\theta}_s^{\lambda},\, h(\bar{\theta}_{\lfloor s\rfloor}^{\lambda}) - H(\bar{\theta}_{\lfloor s\rfloor}^{\lambda}, X_{\lceil s\rceil})\rangle\right]\mathrm{d}s \\ &\quad - 2\lambda\int_{nT}^{t}\mathbb{E}\left[\langle\bar{\zeta}_s^{\lambda,n} - \bar{\theta}_s^{\lambda}, H(\bar{\theta}_{\lfloor s\rfloor}^{\lambda}, X_{\lceil s\rceil}) - H_\lambda(\bar{\theta}_{\lfloor s\rfloor}^{\lambda}, X_{\lceil s\rceil})\rangle\right]\mathrm{d}s. \qquad (53)\end{aligned}$$

Furthermore, by applying Proposition 4 to the first term of (53), and by applying Young's inequality, i.e., $2ab \leq a(2b) \leq L_R a^2/2 + 2b^2/L_R$ for $a, b \geq 0$, to the second and fourth term

of (53), one obtains that

$$\begin{aligned}
\mathbb{E}\left[|\bar{\theta}_t^\lambda - \bar{\zeta}_t^{\lambda,n}|^2\right] &\le 2\lambda L_R \int_{nT}^t \mathbb{E}\left[|\bar{\zeta}_s^{\lambda,n} - \bar{\theta}_s^\lambda|^2\right] \mathrm{d}s \\
&+ \frac{\lambda L_R}{2} \int_{nT}^t \mathbb{E}\left[|\bar{\zeta}_s^{\lambda,n} - \bar{\theta}_s^\lambda|^2\right] \mathrm{d}s + \int_{nT}^t \frac{2\lambda}{L_R} \mathbb{E}\left[|h(\bar{\theta}_s^\lambda) - h(\bar{\theta}_{\lfloor s \rfloor}^\lambda)|^2\right] \mathrm{d}s \\
&+ \int_{nT}^t \Big(- 2\lambda \mathbb{E}\left[\langle \bar{\zeta}_s^{\lambda,n} - \bar{\theta}_s^\lambda,\, h(\bar{\theta}_{\lfloor s \rfloor}^\lambda) - H(\bar{\theta}_{\lfloor s \rfloor}^\lambda, X_{\lceil s \rceil})\rangle\right] \Big) \mathrm{d}s \\
&+ \frac{\lambda L_R}{2} \int_{nT}^t \mathbb{E}\left[|\bar{\zeta}_s^{\lambda,n} - \bar{\theta}_s^\lambda|^2\right] \mathrm{d}s \\
&+ \int_{nT}^t \frac{2\lambda}{L_R} \mathbb{E}\left[|H(\bar{\theta}_{\lfloor s \rfloor}^\lambda, X_{\lceil s \rceil}) - H_{\lambda,c}(\bar{\theta}_{\lfloor s \rfloor}^\lambda, X_{\lceil s \rceil})|^2\right] \mathrm{d}s \\
&= 3\lambda L_R \int_{nT}^t \mathbb{E}\left[|\bar{\zeta}_s^{\lambda,n} - \bar{\theta}_s^\lambda|^2\right] \mathrm{d}s + \int_{nT}^t \left(A_s^{\lambda,n} + B_s^{\lambda,n} + D_s^{\lambda,n}\right) \mathrm{d}s, \qquad (54)
\end{aligned}$$

where

$$\begin{aligned}
A_t^{\lambda,n} &:= \frac{2\lambda}{L_R} \mathbb{E}\left[|h(\bar{\theta}_t^\lambda) - h(\bar{\theta}_{\lfloor t \rfloor}^\lambda)|^2\right] \\
B_t^{\lambda,n} &:= -2\lambda \mathbb{E}\left[\langle \bar{\zeta}_t^{\lambda,n} - \bar{\theta}_t^\lambda,\, h(\bar{\theta}_{\lfloor t \rfloor}^\lambda) - H(\bar{\theta}_{\lfloor t \rfloor}^\lambda, X_{\lceil t \rceil})\rangle\right] \\
D_t^{\lambda,n} &:= \frac{2\lambda}{L_R} \mathbb{E}\left[|H(\bar{\theta}_{\lfloor t \rfloor}^\lambda, X_{\lceil t \rceil}) - H_{\lambda,c}(\bar{\theta}_{\lfloor t \rfloor}^\lambda, X_{\lceil t \rceil})|^2\right].
\end{aligned}$$

In addition, using the definition of $\bar{\theta}_t^\lambda$ and the inequality of (36), we have

$$\begin{aligned}
|\bar{\theta}_t^\lambda - \bar{\theta}_{\lfloor t \rfloor}^\lambda|^4 &\le \left| \lambda \int_{\lfloor t \rfloor}^t |H_{\lambda,c}(\bar{\theta}_{\lfloor s \rfloor}^\lambda, X_{\lceil s \rceil})| ds + \sqrt{2\lambda\beta^{-1}} |B_t^\lambda - B_{\lfloor t \rfloor}^\lambda| \right|^4 \\
&\le 8\lambda^2 \Big(\lambda^2 |H_{\lambda,c}(\bar{\theta}_{\lfloor t \rfloor}^\lambda, X_{\lceil t \rceil})|^4 + 4\beta^{-2} |B_t^\lambda - B_{\lfloor t \rfloor}^\lambda|^4 \Big) \\
&\le 8\lambda^2 \Big((4d(1+\lambda^2) + 2\eta^2 |\bar{\theta}_{\lfloor t \rfloor}^\lambda|^2)^2 + 4\beta^{-2} |B_t^\lambda - B_{\lfloor t \rfloor}^\lambda|^4 \Big) \\
&\le 2^5 \lambda^2 \Big(2^3 d^2 (1+\lambda^4) + \eta^4 |\bar{\theta}_{\lfloor t \rfloor}^\lambda|^4 + \beta^{-2} |B_t^\lambda - B_{\lfloor t \rfloor}^\lambda|^4 \Big),
\end{aligned}$$

which yields that

$$\begin{aligned}
\sqrt{\mathbb{E}|\bar{\theta}_t^\lambda - \bar{\theta}_{\lfloor t \rfloor}^\lambda|^4} &\le 2^{5/2} \sqrt{16d^2 + \eta^4 \mathbb{E}\left[|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^4\right] + \beta^{-2} \mathbb{E}\left[|B_t^\lambda - B_{\lfloor t \rfloor}^\lambda|^4\right]}\, \lambda \\
&\le \widetilde{C}_1 \lambda \qquad (55)
\end{aligned}$$

where Lemma 13 is used for the first inequality, and

$$\widetilde{C}_1 := 2^{5/2} \sqrt{16d^2 + \eta^4 (\mathbb{E}\left[|\theta_0|^4\right] + A_2/\eta^2) + \frac{3}{\beta^2} d^2}.$$

Using Remark 6, Lemma 13 and (55), $A_t^{\lambda,m}$ can be bounded as follows:

$$\begin{aligned} A_t^{\lambda,n} &\leq \frac{2\lambda L_h^2}{L_R}\mathbb{E}\left[(1+|\bar{\theta}_t^\lambda|+|\bar{\theta}_{\lfloor t\rfloor}^\lambda)^{2l}|\bar{\theta}_t^\lambda-\bar{\theta}_{\lfloor t\rfloor}^\lambda|^2\right] \\ &\leq \frac{2\lambda L_h^2}{L_R}\sqrt{\mathbb{E}\left[(1+|\bar{\theta}_t^\lambda|+|\bar{\theta}_{\lfloor t\rfloor}^\lambda)^{4l}\right]}\sqrt{\mathbb{E}\left[|\bar{\theta}_t^\lambda-\bar{\theta}_{\lfloor t\rfloor}^\lambda|^4\right]} \\ &\leq \frac{2\lambda L_h^2}{L_R}3^{2l}\sqrt{(1+\mathbb{E}\left[|\bar{\theta}_t^\lambda|^{4l}\right]+\mathbb{E}\left[|\bar{\theta}_{\lfloor t\rfloor}^\lambda|^{4l}\right])}\sqrt{\mathbb{E}\left[|\bar{\theta}_t^\lambda-\bar{\theta}_{\lfloor t\rfloor}^\lambda|^4\right]} \\ &\leq \bar{C}_1\lambda^2 \end{aligned} \tag{56}$$

where $\bar{C}_1=\frac{2L_h^2 9^l}{L_R}\sqrt{1+2\mathbb{E}\left[|\bar{\theta}_0^\lambda|^{4l}\right]+2\frac{A_{2l}}{\eta^2}}\widetilde{C}_1$. To estimate $B_t^{\lambda,n}$, we observe that

$$\begin{aligned} B_t^{\lambda,n} &= -2\lambda\mathbb{E}\left[\langle\bar{\zeta}_t^{\lambda,n}-\bar{\theta}_{\lfloor t\rfloor}^\lambda,\, h(\bar{\theta}_{\lfloor t\rfloor}^\lambda)-H(\bar{\theta}_{\lfloor t\rfloor}^\lambda,X_{\lceil t\rceil})\rangle\right] \\ &\quad -2\lambda\mathbb{E}\left[\langle\bar{\theta}_{\lfloor t\rfloor}^\lambda-\bar{\theta}_t^\lambda,\, h(\bar{\theta}_{\lfloor t\rfloor}^\lambda)-H(\bar{\theta}_{\lfloor t\rfloor}^\lambda,X_{\lceil t\rceil})\rangle\right] \\ &\leq -2\lambda\mathbb{E}\left[\mathbb{E}\left[\langle\bar{\zeta}_t^{\lambda,n}-\bar{\theta}_{\lfloor t\rfloor}^\lambda,\, h(\bar{\theta}_{\lfloor t\rfloor}^\lambda)-H(\bar{\theta}_{\lfloor t\rfloor}^\lambda,X_{\lceil t\rceil})\rangle\middle|\bar{\zeta}_t^{\lambda,n},\bar{\theta}_{\lfloor t\rfloor}^\lambda\right]\right] \\ &\quad -2\lambda\mathbb{E}\left[\langle\bar{\theta}_{\lfloor t\rfloor}^\lambda-\bar{\theta}_t^\lambda,\, h(\bar{\theta}_{\lfloor t\rfloor}^\lambda)-H(\bar{\theta}_{\lfloor t\rfloor}^\lambda,X_{\lceil t\rceil})\rangle\right] \\ &\leq -2\lambda\mathbb{E}\left[\langle\lambda\int_{\lfloor t\rfloor}^t H_\lambda(\bar{\theta}_{\lfloor s\rfloor}^\lambda,X_{\lceil s\rceil})\mathrm{d}s-\sqrt{\frac{2\lambda}{\beta}}B_{t-\lfloor t\rfloor}^\lambda,\, h(\bar{\theta}_{\lfloor t\rfloor}^\lambda)-H(\bar{\theta}_{\lfloor t\rfloor}^\lambda,X_{\lceil t\rceil})\rangle\right], \end{aligned}$$

where we have used that H is an unbiased estimator of h, i.e., $H(\theta,X_0)=h(x)$ for all $x\in\mathbb{R}^m$, to obtain the last inequality. Moreover, using Remark 22 and Lemma 13, we have

$$\begin{aligned} B_t^{\lambda,n} &\leq -2\lambda^2\mathbb{E}\left[\langle H_\lambda(\bar{\theta}_{\lfloor t\rfloor}^\lambda,X_{\lceil t\rceil}),\, h(\bar{\theta}_{\lfloor t\rfloor}^\lambda)-H(\bar{\theta}_{\lfloor t\rfloor}^\lambda,X_{\lceil t\rceil})\rangle\right] \\ &\leq 2\lambda^2\mathbb{E}\left[\langle|H_\lambda(\bar{\theta}_{\lfloor t\rfloor}^\lambda,X_{\lceil t\rceil})|,\, |h(\bar{\theta}_{\lfloor t\rfloor}^\lambda)|+|H(\bar{\theta}_{\lfloor t\rfloor}^\lambda,X_{\lceil t\rceil})|\rangle\right] \\ &\leq 4\lambda^2\mathbb{E}\left[|H_\lambda(\bar{\theta}_{\lfloor t\rfloor}^\lambda,X_{\lceil t\rceil})|^2\right] \\ &\leq 12\left(\mathbb{E}\left[|K_G(X_0)|^2(1+\mathbb{E}\left[|\bar{\theta}_{\lfloor t\rfloor}^\lambda|^{2q}\right])\right]+d+\eta^2\mathbb{E}\left[|\bar{\theta}_{\lfloor t\rfloor}^\lambda|^{4r+2}\right]\right)\lambda^2 \\ &\leq \bar{C}_2\lambda^2,, \end{aligned} \tag{57}$$

where the constant $\bar{C}_2$ is given by

$$\bar{C}_2=12\sqrt{\mathbb{E}\left[|K_G(X_0)|^2\right]\left(1+\mathbb{E}\left[|\theta_0|^{2q}\right]+\frac{A_q}{\eta^2}\right)+d+\eta^2\left(\mathbb{E}\left[|\theta_0|^{4r+2}\right]+\frac{A_{2r+1}}{\eta^2}\right)}.$$

Moreover, $D_t^{\lambda,n}$ can be estimated as follows, from Remark 21 and Lemma 13,

$$\begin{aligned} D_t^{\lambda,n} &\leq \frac{18\lambda^2}{L_R}\left[8\mathbb{E}\left[|K_G(X_0)|^4\right](1+\mathbb{E}\left[|\bar{\theta}_{\lfloor t\rfloor}^\lambda|^{4q}\right])+d+\eta^2\mathbb{E}\left[|\bar{\theta}_{\lfloor t\rfloor}^\lambda|^{4r+2}\right]\right] \\ &\leq \bar{C}_3\lambda^2, \end{aligned} \tag{58}$$

where the independence of $\bar{\theta}^{\lambda}_{\lfloor s \rfloor}$ and $X_{\lceil s \rceil}$ is used, and $\bar{C}_3$ is given by

$$\bar{C}_3 = \frac{18}{L_R}\left[8\mathbb{E}\left[|K_G(X_0)|^4\right](1+\mathbb{E}\left[|\bar{\theta}^{\lambda}_0|^{4q}\right]+A_{2q}/\eta^2)+d+\eta^2(\mathbb{E}\left[|\bar{\theta}^{\lambda}_0|^{4r+2}\right]+A_{2r+1}/\eta^2)\right].$$

Substituting (56), (57), and (58) into (53), one can derive

$$\begin{aligned}\mathbb{E}\left[|\bar{\theta}^{\lambda}_t-\bar{\zeta}^{\lambda,n}_t|^2\right] &\leq 3\lambda L_R\int_{nT}^{t}\mathbb{E}\left[|\bar{\theta}^{\lambda}_s-\bar{\zeta}^{\lambda,n}_s|^2\right]\mathrm{d}s+\int_{nT}^{t}(\bar{C}_1+\bar{C}_2+\bar{C}_3)\lambda^2\mathrm{d}s\\ &\leq 3\lambda L_R\int_{nT}^{t}\mathbb{E}\left[|\bar{\theta}^{\lambda}_s-\bar{\zeta}^{\lambda,n}_s|^2\right]\mathrm{d}s+(\bar{C}_1+\bar{C}_2+\bar{C}_3)\lambda<\infty\end{aligned}$$

where the second inequality follows from the fact that $(t-nT)\leq T\leq\frac{1}{\lambda}$ and the use of Gronwall's inequality gives

$$\mathbb{E}|\bar{\theta}^{\lambda}_t-\bar{\zeta}^{\lambda,n}_t|^2\leq e^{3L_R}(\bar{C}_1+\bar{C}_2+\bar{C}_3)\lambda.$$

□

Proof of Lemma 19. Recall that $Z^{\lambda}_t=\bar{\zeta}^{\lambda,0}_t$. For $t\in(nT,(n+1)T]$, $n\in\mathbb{N}_0$, we can write

$$\begin{aligned}W_1\left(\mathcal{L}(\bar{\zeta}^{\lambda,n}_t),\mathcal{L}(Z^{\lambda}_t)\right) &\leq \sum_{k=1}^{n}W_1\left(\mathcal{L}(\bar{\zeta}^{\lambda,k}_t),\mathcal{L}(\bar{\zeta}^{\lambda,k-1}_t)\right)\\ &\leq \sum_{k=1}^{n}w_{1,2}\left(\mathcal{L}(\bar{\zeta}^{\lambda,k}_t),\mathcal{L}(\bar{\zeta}^{\lambda,k-1}_t)\right),\end{aligned}\tag{59}$$

where we have used the fact $W_1(\mu,\nu)\leq w_{1,2}(\mu,\nu)$ for $\mu,\nu\in\mathcal{P}_{V_2}(\mathbb{R}^d)$ for the second inequality. Using Lemma 17 and $\lambda(t-kT)\geq n-k$, we further calculate

$$\begin{aligned}&w_{1,2}\left(\mathcal{L}(\bar{\zeta}^{\lambda,k}_t),\mathcal{L}(\bar{\zeta}^{\lambda,k-1}_t)\right)\\ &\leq \hat{c}e^{-C_0\lambda(t-kT)}w_{1,2}\left(\mathcal{L}(\bar{\theta}^{\lambda}_{kT}),\mathcal{L}(\bar{\zeta}^{\lambda,k-1}_{kT})\right)\\ &\leq \hat{c}e^{-C_0(n-k)}w_{1,2}\left(\mathcal{L}(\bar{\theta}^{\lambda}_{kT}),\mathcal{L}(\bar{\zeta}^{\lambda,k-1}_{kT})\right)\\ &\leq \hat{c}e^{-C_0(n-k)}W_2\left(\mathcal{L}(\bar{\theta}^{\lambda}_{kT}),\mathcal{L}(\bar{\zeta}^{\lambda,k-1}_{kT})\right)\sqrt{\mathbb{E}\left[\left|1+V_2(\bar{\theta}^{\lambda}_{kT})+V_2(\bar{\zeta}^{\lambda,k-1}_{kT})\right|^2\right]}\\ &\leq \hat{c}e^{-C_0(n-k)}W_2\left(\mathcal{L}(\bar{\theta}^{\lambda}_{kT}),\mathcal{L}(\bar{\zeta}^{\lambda,k-1}_{kT})\right)\left(1+\sqrt{\mathbb{E}[V_4(\bar{\theta}^{\lambda}_{kT})]}+\sqrt{\mathbb{E}[V_4(\bar{\zeta}^{\lambda,k-1}_{kT})]}\right)\\ &\leq \sqrt{\lambda}\hat{c}e^{-C_0(n-k)}\sqrt{e^{3L_R}(\bar{C}_1+\bar{C}_2+\bar{C}_3)}\Bigg(1+\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}}\\ &+\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}+\frac{\tilde{c}(4)}{\bar{c}(4)}}\Bigg)\end{aligned}\tag{60}$$

where Lemma 15, 18 and 16 are used for the last inequality. By substituting (60) into (59), we obtain

$$\begin{aligned} W_1\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,n}\right),\mathcal{L}\left(Z_t^{\lambda}\right)\right) &\leq \sqrt{\lambda}\hat{c}\sqrt{e^{3L_R}(\bar{C}_1+\bar{C}_2+\bar{C}_3)}\Bigg[1+\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}} \\ &+ \sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}+\frac{\tilde{c}(4)}{\bar{c}(4)}}\Bigg]\sum_{k=1}^{n}e^{-C_0(n-k)} \\ &\leq z_1\sqrt{\lambda}, \end{aligned}$$

where

$$\begin{aligned} z_1 &= \frac{\hat{c}}{1-\exp(-C_0)}\sqrt{e^{3L_R}(\bar{C}_1+\bar{C}_2+\bar{C}_3)} \\ &\times \left[1+\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}}+\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}+\frac{\tilde{c}(4)}{\bar{c}(4)}}\right]. \end{aligned}$$

□

Proof of Lemma 20. We begin by observing that, for $t\in(nT,(n+1)T]$, $n\in\mathbb{N}_0$,

$$\begin{aligned} W_2\left(\mathcal{L}(\bar{\zeta}_t^{\lambda,k}),\mathcal{L}(\bar{\zeta}_t^{\lambda,k-1})\right) &\leq \sqrt{2w_{1,2}\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,k}\right),\mathcal{L}\left(\bar{\zeta}_t^{\lambda,k-1}\right)\right)} \\ &\leq \lambda^{1/4}e^{-C_0(n-k)/2}\Bigg[\hat{c}\sqrt{e^{3L_R}(\bar{C}_1+\bar{C}_2+\bar{C}_3)} \\ &\times \Bigg(1+\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}} \\ &+ \sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}+\frac{\tilde{c}(4)}{\bar{c}(4)}}\Bigg)\Bigg]^{1/2}, \end{aligned}$$

where the first inequality holds due to (19), and the second inequality follows from (60). Consequently, we derive

$$\begin{aligned} W_2\left(\mathcal{L}(\bar{\zeta}_t^{\lambda,n}),\mathcal{L}(Z_t^{\lambda})\right) &\leq \sum_{k=1}^{n}W_2\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,k}\right),\mathcal{L}(\bar{\zeta}_t^{\lambda,k-1})\right) \\ &\leq \lambda^{1/4}\Bigg[\hat{c}\sqrt{e^{3L_R}(\bar{C}_1+\bar{C}_2+\bar{C}_3)}\Bigg(1+\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}} \\ &+ \sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}+\frac{\tilde{c}(4)}{\bar{c}(4)}}\Bigg)\Bigg]^{1/2}\sum_{k=1}^{n}e^{-C_0(n-k)/2} \\ &\leq \lambda^{1/4}z_2 \end{aligned}$$

where

$$z_2=\frac{\sqrt{(1-\exp(-C_0))z_1}}{1-\exp(-C_0/2)}.$$

□

C. Table of Constants

Table 9 displays full expressions for constants which appear in the main results of this paper. In addition, Table 10 shows all main constants and their dependency on key parameters such as d, β, and the moments of $K_G(X_0)$.

References

S. Ahn, A. Korattikara, and M. Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. *International Conference on Learning Representations*, 2021.

L. Aitchison. A statistical theory of cold posteriors in deep neural networks. *International Conference on Learning Representations*, 2021.

L. Balles and P. Hennig. The sign, magnitude and variance of stochastic gradients. *International Conference on Machine Learning*, 2018.

M. Barkhagen, N. H. Chau, É. Moulines, M. Rásonyi, S. Sabanis, and Y. Zhang. On stochastic gradient Langevin dynamics with dependent data streams in the logconcave case. *Bernoulli*, 27(1):1–33, 2021.

Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

N. Brosse, E. Moulines, and A. Durmus. The promises and pitfalls of stochastic gradient langevin dynamics. *Advances in Neural Information Processing Systems*, 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.

H. N. 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.

T. Chen, E. B. Fox, and C. Guestrin. Stochastic gradient Hamiltoniam monte carlo. *International Conference on Machine Learning*, 2014.

X. Chen, S. Liu, R. Sun, and M. Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. *International Conference on Learning Representations*, 2019.

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.

W. Deng, Q. Feng, L. Gao, and G. Lin. Non-convex learning via replica exchange stochastic gradient MCMC. *International Conference on Machine Learning*, 2020a.

W. Deng, G. Lin, and F. Liang. A contour stochastic gradient Langevin dynamics algorithm for simulations of multi-modal distributions. *Conference on Neural Information Processing Systems*, 2020b.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12, 2011.

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

A. Eberle, A. Guillin, and R. Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *Annals of Probability*, pages 1982–2010, 2019.

M. A. Erdogdu, L. Mackey, and O. Shamir. Global non-convex optimization with discretized diffusions. *Conference on Neural Information Processing Systems*, 2018.

B. Heo, S. Chun, S. Oh, D. H. S. Yun, G. Kim, Y. Uh, and J. Ha. Adamp: slowing down the slodown for momentum optimizers on scale-invariant weights. *International Conference on Learning Representations*, 2021.

G. Huang, Z. Liu, L. Maaten, and K. Weinberger. Densely connected convolutional networks. *IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

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

C. Hwang. Laplace's method revisited: weak convergence of probability measures. *The Annals of Probability*, pages 1177–1182, 1980.

S. Ioffe and C. Szegedy. Deep residual learning for image recognition. *IEEE conference on computer vision and pattern recognition*, pages 248–255, 2016.

N. Keskar and R. Socher. Improving generalization performance by switching from adam and sgd. *arXiv:1712.07628, 2017*, 2017.

D. Kingma and J. Ba. ADAM: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

A. Krizhevsk. Learning multiple layers of features from tiny images. 2009. URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

A. Krizhevsky, V. Nair, and G. Hinton. Cifar-10 (canadian institute for advanced research). URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

N. V. Krylov. Extremal properties of the solutions of stochastic equations. *Theory of Probability and its Applications,*, 29(2):205–217, 1985.

N. V. Krylov. A simple proof of the existence of a solution to the Itô's equation with monotone coefficients. *Theory of Probability and its Applications,*, 35(3):583–587, 1990.

H. Lee, A. Riesteski, and R. Ge. Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering Langevin monte carlo. *Advances in Neural Information Processing Systems*, 2018.

L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. On the variance of the adaptive learning rate and beyond. *International Conference on Learning Representations*, 2020.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019.

A. Lovas, I. Lytas, M. Rasonyi, and S. Sabanis. Taming neural networks with tusla: Non-convex learning via adaptive stochastic gradient langevin algorithms. *arXiv preprint arXiv:2006.14514*, 2020.

L. Luo, Y. Xiong, Y. Liu, and X. Sun. Adaptive gradient methods with dynamic bound of learning rate. *International Conference on Learning Representations*, 2019.

O. Mangoubi and N. K. Vishnoi. Convex optimization with unbounded nonconvex orcles using simulated annealing. *Conference on Learning Theory*, 2018.

M. Marcus, B. Santorini, M. Marcinkiewicz, and A. Taylor. Treebank-3. 1999. URL `https://doi.org/10.35111/gq1x-j780`.

S. Merity, N. S. Keskar, and R. Socher. Regularizing and optimizing lstm language models. *International Conference on Learning Representations*, 2018.

R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. *Proceedings of the 30th International Conference on Machine Learning*, 2013.

B. Polyak and A. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal oon control and optimization*, 30:835–855, 1992.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. *Conference on Learning Theory*, 2017.

S. Reddi, S. Kale, and S. Kumar. On the convergence of ADAM and beyond. *International Conference on Learning Representations*, 2018.

G. O. Roberts and R. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

S. Sabanis. A note on tamed euler approximations. *Electronic Communications in Probability*, 18(47):1–10, 2013.

S. Sabanis. Euler approximation with varying coefficients: the case of superlienarly growing diffusion coefficients. *Annals of Applied Probability*, 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.

K. Simonya and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015.

T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 2012.

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.

F. Wenzel, K. Roth, B. Veeling, J. Swiatkowski, L. Tran, S. Mandt, J. Snoek, T. Salimans, R. Jenatton, and S. Nowzin. How good is the bayes posterior in deep neural networks really? *International Conference on Machine Learning*, 2020.

A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht. The marginal value of adaptive gradient methods in machine learning. *Advances in Neural Information Processing Systems*, 2017.

P. Xu, J. Chen, D. Zou, and Q. Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. *Conference on Neural Information Processing Systems*, 2018.

M. Zaheer, S. Reddi, D. Sachan, S. Kale, and S. Kumar. Adaptive methods for nonconvex optimization. *Advances in Neural Information Processing Systems*, 2018.

R. Zhang, C. Li, J. Zhang, C. Chen, and A. Wilson. Cyclical stochastic gradient MCMC for bayesian deep learning. *International Conference on Learning Representations*, 2020.

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.

J. Zhuang, T. Tang, Y. Ding, S. Tatikonda, N. Dvornek, X. Papademetris, and J. Duncan. Adabelief optimizer: adapting stepsizes by the belief in observed gradients. *Advances in Neural Information Processing Systems*, 2020.

Table 9: Explicit expressions for main constants.

Symbol	Full Expression
M	$\max\left\{M_0, 1, \frac{4d}{(2-\eta)\eta}, \frac{2\sqrt{d}}{\eta(2-\eta)}, \frac{2^{2p-2}p(2p-1)d}{\eta\beta}\right\}$
$\widetilde{D}_k$	$2^{k-1}\Big[(2\sqrt{M})^k(\mathbb{E}[K_G(X_0)](1+M^q)+d+2\eta M^{2r+1})^{k/2} +(8d+2\eta^2M^2)^k\Big], \quad k=1,\cdots,8(2r+1)$
A_p	$\eta^2M^{2p}+M^{2p}\sum_{k=1}^{p}\binom{p}{k}\widetilde{D}_k +2^{2p-3}p(2p-1)\Big(\frac{2dM^{2p-2}}{\beta}\sum_{k=0}^{p-1}\binom{p}{k}\widetilde{D}_k+\frac{2}{\beta}\Big(\frac{2}{\beta}\Big)^{p-1}d^p(2p-1)!!\Big),$ for $p=1,\cdots,8(2r+1)$
$\overline{M}_p$	$\sqrt{\frac{1}{3}+4B/(3A)+4d/(3A\beta)+4(p-2)/(3A\beta)}$
$\bar{c}(p)$	$\frac{Ap}{4}, \quad p=1,\cdots,8(2r+1)$
$\tilde{c}(p)$	$\frac{3}{4}Apv_p(\overline{M}_p), \quad p=1,\cdots,8(2r+1)$
$\bar{C}_1$	$\frac{L^2 2^{2\rho+5/2}3^{2l}}{L_R}(1+\mathbb{E}\lvert X_0\rvert^{2\rho})\sqrt{(1+2\mathbb{E}\lvert\bar{\theta}_0^{\lambda}\rvert^{4l}+2\frac{A_{2l}}{\eta^2})} \times\sqrt{16d^2+\eta^4(\mathbb{E}\lvert\theta_0\rvert^4+A_2/\eta^2)+\frac{3}{\beta^2}d^2}$
$\bar{C}_2$	$12\sqrt{\mathbb{E}\lvert K_G(X_0)\rvert^2\left(1+\mathbb{E}\lvert\theta_0\rvert^{2q}+\frac{A_q}{\eta^2}\right)+d+\eta^2\left(\mathbb{E}\lvert\theta_0\rvert^{4r+2}+\frac{A_{2r+1}}{\eta^2}\right)}$
$\bar{C}_3$	$\frac{18}{L_R}\Big[8\mathbb{E}\lvert K_G(X_0)\rvert^4(1+\mathbb{E}\lvert\bar{\theta}_0^{\lambda}\rvert^{4q}+A_{2q}/\eta^2)+d+\eta^2(\mathbb{E}\lvert\bar{\theta}_0^{\lambda}\rvert^{4r+2}+A_{2r+1}/\eta^2)\Big].$
z_1	$\frac{\hat{c}\sqrt{e^{3L_R}(\bar{C}_1+\bar{C}_2+\bar{C}_3)}}{1-\exp(-C_0)}\Big[1+\sqrt{2\mathbb{E}\lvert\theta_0\rvert^4+2+2\frac{A_2}{\eta^2}}+\sqrt{2\mathbb{E}\lvert\theta_0\rvert^4+2+2\frac{A_2}{\eta^2}+\frac{\tilde{c}(4)}{\bar{c}(4)}}\Big]$
z_2	$\frac{\sqrt{(1-\exp(-C_0))z_1}}{1-\exp(-C_0/2)}$
R_0	$\max\{\sqrt{B/A}, \sqrt{2d/(\beta L\mathbb{E}(1+\lvert X_0\rvert)^{\rho})}\}$
$\hat{c}$	See Lemma 17.
C_0	See Lemma 17.
C_1	$(z_1+\sqrt{e^{3L_R}(\bar{C}_1+\bar{C}_2+\bar{C}_3)})$
C_2	$\hat{c}\Big(1+\mathbb{E}[V_2(\theta_0)]+\int_{\mathbb{R}^d}V_2(\theta)\pi_{\beta}(d\theta)\Big)$
C_3	$\sqrt{e^{3a}(\bar{C}_1+\bar{C}_2+\bar{C}_3)}+z_2$
C_4	$\sqrt{2\hat{c}\left(1+\mathbb{E}\left[V_2\left(\theta_0\right)\right]+\int_{\mathbb{R}^d}V_2(\theta)\pi_{\beta}(d\theta)\right)}$
C_5	$\frac{C_0}{2}$
C_6	$\left(r_1+\frac{r_1 2^{2r}}{2r+2}\sqrt{\mathbb{E}\lvert\theta_0^{\lambda}\rvert^{4r+2}+\frac{A_{2r+1}}{\eta^2}}+\frac{r_1 2^{2r}}{2r+2}\sqrt{\mathbb{E}\lvert Z_{\infty}\rvert^{4r+2}}\right)$
A	$2^q\mathbb{E}[1+K_G(X_0)]$
B	$3(2^{q+1}\mathbb{E}[1+K_G(X_0)])^{q+2}/\eta^{q+1}$
K	$L\mathbb{E}[(1+\lvert X_0\rvert)^{\rho}](1+4R_0)^{2r+1}$

Table 10: Main constants and their dependency to key parameters

Constant	Key parameters		
	d	β	Moments of X_0
A	-	-	$\mathcal{O}(\mathbb{E}[K_G(X_0)])$
B	-	-	$\mathcal{O}(\mathbb{E}[K_G(X_0)^{q+2}])$
R	-	-	$\mathcal{O}(\mathbb{E}[\lvert X_0\rvert^{\rho}])$
a	-	-	$\mathcal{O}\left(\mathbb{E}[\lvert X_0\rvert^{\rho(q-1)}]\right)$
A_p	$poly(d)$	$\mathcal{O}\left(\frac{d}{\beta}\right)$	$\mathcal{O}\left(\mathbb{E}[K_G(X_0)^{\frac{p}{2}}]\right)$
C_0	$poly\left(\frac{d}{\beta}\right)$	$poly\left(\frac{d}{\beta}\right)$	$\mathbb{E}\left[poly(K_G(X_0)^{\frac{q+1}{2}}\right]$