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

Anonymous authors
Paper under double-blind review

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 neural networks. 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.

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 non-convex 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 non-convex 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. 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 & Hinton (2012)), Adam (Kingma & Ba (2015)) have been developed to tackle this problem and improve training speed by diagonally scaling the gradient by some function of the past gradients.

A family of Langevin based algorithms has been another important stream of literature on the stochastic optimization. They are built on the theoretical fact that the Langevin stochastic differential equation, (6), converges to its unique invariant measure, which concentrates on the global minimizers of the objective function as $\beta \to \infty$, see Hwang (1980). Since the convergence property remains true for nonconvex optimization problems, the global convergence of the stochastic gradient Langevin dynamics (SGLD) and its variants has been extensively studied in a nonconvex setting (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 statistics and Bayesian learning (Roberts & Tweedie, 1996; Durmus & Moulines, 2017; Dalalyan, 2017; Brosse et al., 2019; Welling & Teh, 2011; Deng et al., 2020a;b).

Motivated by the aforementioned developments in the field, we propose a new class of Langevin algorithms which is based on recent advances of Euler's polygonal approximations for Langevin SDEs. The idea of this new form of Euler's polygonal approximations for SDEs with monotone coefficients originates from the articles Krylov (1985) and Krylov (1990). We name this new class as *polygonal unadjusted Langevin algorithms*. Moreover, it is versatile enough to incorporate further features to address other known shortcomings of adaptive optimizers. 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

$$\theta_0^\lambda := \theta_0, \qquad \theta_{n+1}^\lambda := \theta_n^\lambda - \lambda G_\lambda(\theta_n^\lambda, X_{n+1}) + \sqrt{2\lambda\beta^{-1}}\xi_{n+1}, \quad n \in \mathbb{N}, \tag{1}$$

where 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 $G_\lambda : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ satisfies the following three properties:

1. For every $\lambda > 0$, There exist constants $K_\lambda > 0$ and $\rho_1 \geq 0$ such that $|G_\lambda(\theta, x)| \leq K_\lambda(1 + |x|)^{\rho_1}(1 + |\theta|)$ for every $\theta \in \mathbb{R}^d$ and $x \in \mathbb{R}^m$.
2. There exist constants $\gamma \geq 1/2$, $K_2 > 0$ and $\rho_2,\ \rho_3 \geq 0$ such that for all $\lambda > 0$,
$$|G_\lambda(\theta, x) - G(\theta, x)| \leq \lambda^\gamma K_2(1 + |x|)^{\rho_2}(1 + |\theta|)^{\rho_3}$$
for every $\theta \in \mathbb{R}^d$ and $x \in \mathbb{R}^m$, where G is the (unbiased) stochastic gradient of the objective function of the optimization problem under study.
3. There exist constants λ_{max} and $\delta \in \{1,\ 2\}$ such that for any $\lambda \leq \lambda_{max}$,
$$\liminf_{|\theta|\to\infty} \mathbb{E}\left[\langle \frac{\theta}{|\theta|^\delta}, G_\lambda(\theta, X_0)\rangle - \frac{2\lambda}{|\theta|^\delta}|G_\lambda(\theta, X_0)|^2\right] > 0.$$

One obtains our new algorithm THεO POULA by considering the case where $G_\lambda(\theta, x)$ is the vector with entries $H_{\lambda,c}^{(i)}(\theta, x)$ as given by (8), for $i \in \{1, \ldots, d\}$. Its name is formed from its description, namely Tamed Hybrid ε-Order POlygonal Unadjusted Langevin Algorithm and its full detailed analysis (including its convergence properties) are given in Section 3. We note that THεO POULA 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).

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

Most research on Langevin based algorithms 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 GLD, SGLD and SVRG-LD. The incorporation of dependent data streams in the analysis of SGLD algorithms has been achieved in Barkhagen et al. (2021) and in Chau et al. (2019), and local conditions have been studied in Zhang et al. (2019). Recently, TUSLA of 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. See Hutzenthaler et al. (2012) and Sabanis (2013) for the rationale of taming techniques. 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 other popular adaptive gradient methods. We refer to Appendix F.3 for the reader who is interested in recent progress on sampling and Bayesian neural networks.

Adaptive learning rate methods such as AdaGrad (Duchi et al. (2011)), RMSProp (Tieleman & Hinton (2012)) and Adam (Kingma & Ba (2015)) have been successfully applied to neural network models due to their fast training speed. Since 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. 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 this 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, the convergence analysis of these (and other) adaptive learning rate methods is still restrictive since it is only guaranteed to converge to a stationary point (which can be a local minimum or a saddle point) under strong assumptions. Namely, the stochastic gradient is globally Lipschitz continuous and bounded. Note though 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 proposed algorithm, THεO POULA, tries to combines both advantages: namely, 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 outperform popular stochastic optimization methods such as SGD, Adam, AMSGrad, RMSProp, AdaBound and AdaBelief for deep learning tasks. The major strengths of our work over related algorithms are summarized as follows:

- **(Global convergence)** We provide a global convergence analysis of THεO POULA for nonconvex optimization where the stochastic gradient of the objective is locally Lipscthiz continuous. Moreover, non-asymptotic estimates for the expected excess risk are derived.
- **(Stable and fast training)** THεO POULA achieves a stable and fast training process using the (element-wise) taming technique, (element-wise) boosting function and averaging, which are theoretically well-designed. Furthermore, we validate the effectiveness of the taming and boosting functions through several empirical experiments.
- **(Good generalization)** While THεO POULA behaves like adaptive learning rate methods in the early training phase, it takes an almost global learning rate near an optimal point. That is, THεO POULA 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 THεO POULA outperforms the other optimization methods in *generalization* measured by test accuracy for various deep learning tasks.

2 Motivating Example

The local Lipschitz continuity of gradients and its effect on the performance of optimization methods are relatively under-studied. Most relevant studies assume that the stochastic gradient is global Lipscthiz continuous and bounded (Kingma & Ba, 2015; Xu et al., 2018; Brosse et al., 2018; Duchi et al., 2011; Tieleman & Hinton, 2012; Reddi et al., 2018; Chen et al., 2019; Liu et al., 2020; Luo et al., 2019; Zhuang et al., 2020) although it is not true for neural network problems. This section provides a simple, one-dimensional optimization problem that illustrates the convergence issue of popular stochastic gradient methods when the stochastic gradient is locally Lipschitz continuous, i.e., the gradient can be super-linearly growing [1].

Consider the following optimization problem:

$$\min_{\theta} u(\theta) = \min_{\theta} \mathbb{E}[U(\theta, X)], \tag{2}$$

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

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

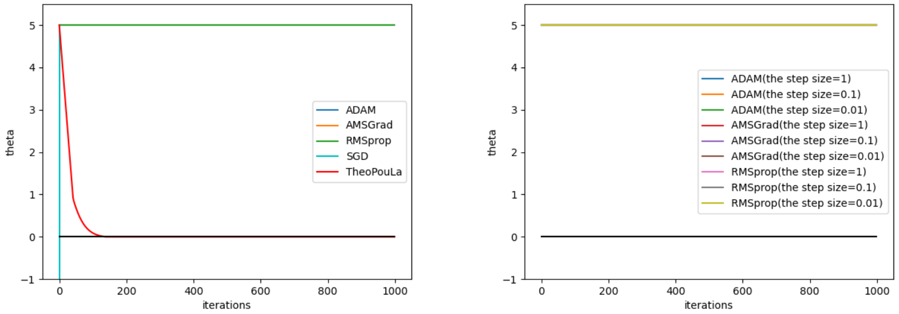
and X is uniformly distributed over $(-2, 2)$, that is, $f_X(x) = \frac{1}{4}\mathbb{1}_{|x| \leq 2}$. Furthermore, the stochastic gradient $G : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is given by

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

where $sgn(\cdot)$ is the sign function. Note that the stochastic gradient G is locally Lipschitz continuous, which satisfies

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

[1] Lovas et al. (2020) used a similar example to show the stability of TUSLA with a different taming function.



(a) default settings

(b) different step sizes

Figure 1: Performance of SGD, Adam, AMSGrad, RMSProp and THεO POULA on an artificial example with the initial value $\theta_0 = 5.0$

for all $x \in \mathbb{R}$ and $\theta, \theta' \in \mathbb{R}$. Also, the optimal value is attained at $\theta = 0$. See Appendix A for more details. Following Reddi et al. (2018), adaptive stochastic gradient methods can be generally written as follows, for $n \in \mathbb{N}$,

$$\begin{aligned} m_n &= \phi_n(G_1, \cdots, G_n), \\ V_n &= \psi_n(G_1, \cdots, G_n), \\ \theta_{n+1} &= \theta_n - \lambda_n \frac{m_n}{\varepsilon + \sqrt{V}_n} \end{aligned} \tag{3}$$

where $G_i := G(\theta_i, X_i)$ is the stochastic gradient evaluated at the i-th iteration, λ_n is the step size and all operations are applied element-wise. Table 1 provides the details for some of the most popular stochastic optimization methods with corresponding averaging functions ϕ_n and ψ_n.

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 $v_n = (1 - \beta_2)v_{n-1} + \beta_2 G_n^2$.

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

We use SGD, Adam, AMSGrad and RMSprop to solve the optimization problem with initial value $\theta_0 = 5$. For hyperparameters of optimization algorithms, we use their default settings provided in PyTorch. Figure 1(a) shows the trajectories of approximate solutions generated by each optimizer. While SGD, Adam, AMSGrad and RMSProp fail to converge to the optimal solution 0, the proposed algorithm, THεO POULA, finds the optimal solution with a reasonable step size, say, 0.01.

Intuitively, the undesirable phenomenon occurs because, in the iterating rule (3), the denominator $\sqrt{V}_n$ excessively dominates the numerator m_n, causing the vanishing gradient problem in the presence of the superlinear gradient. On the contrary, SGD suffers from the exploding gradient problem. Moreover, Figure 1(b) highlights that the problematic behavior cannot be simply resolved by adjusting the learning rate within the Adam-type framework, while THεO POULA perform extremely well even in the presence of such violent non-linearities.

3 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 (stochastic) gradients of neural networks. In other words, the novelty of our algorithm is to utilize a taming function and a boosting function, rather than designing a new V_n as in Adam-type optimizers.

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

3.1 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 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

$$W_p(\mu, \nu) := \inf_{\zeta \in \mathcal{C}(\mu,\nu)} \left(\int_{\mathbb{R}^k} \int_{\mathbb{R}^k} |\theta - \theta'|^p \zeta(\mathrm{d}\theta \mathrm{d}\theta') \right)^{1/p}, \ \mu, \nu \in \mathcal{P}(\mathbb{R}^k). \tag{4}$$

Let $(X_n)_{n \in \mathbb{N}}$ be a sequence of i.i.d. $\mathbb{R}^m$-valued random variables generating the filtration $(\mathcal{G}_n)_{n \in \mathbb{N}}$ and $(\xi_n)_{n \in \mathbb{N}}$ be an $\mathbb{R}^d$-valued Gaussian process with independent components.

Let $F : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ be continuously differentiable function such that $\mathbb{E}[F(\theta, X_0)] < \infty$ for any $\theta \in \mathbb{R}^d$. We consider the following optimization problem

$$\min_\theta u(\theta) \quad = \quad \min_\theta \left(\mathbb{E}[F(\theta, X_0)] + \frac{\eta}{2(r+1)} |\theta|^{2(r+1)} \right) \tag{5}$$

where $\theta \in \mathbb{R}^d$, $\eta \in (0, 1)$ is the regularization parameter and $r \geq \frac{q}{2} + 1$. 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 neural network's parameters. Note that the regularization term, $\frac{\eta}{2(r+1)} |\theta|^{2(r+1)}$, is added in order to guarantee that the dissipativity property holds, since it is essential for the convergence analysis.

Remark 3.1. *For the reader who prefers to consider the optimization problem without the regularization term, i.e. with $\eta = 0$, the dissipative condition (B.1) 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.*

In particular, r depends on the neural network's structure, whereas q is described in Assumption 3.1. Consequently, the stochastic gradient with the regularization term is given by

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

where $G(\theta, x) := \nabla_\theta F(\theta, x)$ for all $x \in \mathbb{R}^m$, $\theta \in \mathbb{R}^d$ and $\eta = 0$ if dissipativity holds for G.

We introduce our main assumptions. The first requirement is that G is locally Lipschitz continuous.

Assumption 3.1. *There exists positive constant L_1, ρ and $q \geq 1$ such that*

$$|G(\theta, x) - G(\theta', x)| \leq L_1(1 + |x|)^\rho (1 + |\theta| + |\theta'|)^{q-1} |\theta - \theta'|.$$

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

Further, conditions on the initial value θ_0 and data process $(X_n)_{n \in \mathbb{N}}$ are imposed as it is common to use weight initialization using the uniform or normal distribution, Assumption 3.2 is mild.

Assumption 3.2. *The process $(X_n)_{n \in \mathbb{N}}$ is a sequence of i.i.d. random variables with $\mathbb{E}[|X_0|^{16\rho(2r+1)}] < \infty$ where ρ is given in Assumption 3.1. In addition, the initial condition is such that $\mathbb{E}[|\theta_0|^{16(2r+1)}] < \infty$.*

We refer to Appendix B for further remarks and key observations regarding the consequences of Assumptions 3.1 and 3.2. We conduct the convergence analysis of THεO POULA by employing elements of the theory of Langevin SDEs. It is shown that, under mild conditions (satisfied by Assumptions 3.1 and 3.2), the so-called (overdamped) Langevin SDE, which is given by

$$\mathrm{d}Z_t = -h(Z_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B_t, \quad t > 0, \tag{6}$$

where $h = \nabla u$ with a (possibly random) initial condition θ_0 and with $(B_t)_{t\geq 0}$ denoting a d-dimensional Brownian motion, admits a unique invariant measure $\pi_\beta(dz) \propto \exp(-\beta u(z))$. Thus, for a sufficiently large β, π_β concentrates around the minimizers of (5).

3.2 Mechanism of THεO POULA

We introduce the mechanism of THεO POULA, which iterately updates as follows:

$$\theta_0^\lambda := \theta_0, \qquad \theta_{n+1}^\lambda := \theta_n^\lambda - \lambda H_{\lambda,c}(\theta_n^\lambda, X_{n+1}) + \sqrt{2\lambda\beta^{-1}}\xi_{n+1}, \quad n \in \mathbb{N}, \tag{7}$$

where $H_{\lambda,c} := (H_{\lambda,c}^{(1)}(\theta,x), \cdots, H_{\lambda,c}^{(d)}(\theta,x))^T$ is given by

$$H_{\lambda,c}^{(i)}(\theta,x) = \underbrace{\frac{G^{(i)}(\theta,x)}{1+\sqrt{\lambda}|G^{(i)}(\theta,x)|}}_{\text{taming function}} \underbrace{\left(1+\frac{\sqrt{\lambda}}{\varepsilon+|G^{(i)}(\theta,x)|}\right)}_{\text{boosting function}} + \underbrace{\eta\frac{\theta^{(i)}|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}}}_{\text{regularization term}}, \tag{8}$$

and $\{\xi_n\}_{n\geq 1}$ is a sequence of independent standard d-dimensional Gaussian random variables. Note that the taming and boosting functions are defined in (8).

THεO POULA 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 improve the performance of the algorithm, and to handle the exploding and vanishing gradient problems of neural networks. For simplicity, we omit the regularization term, that is, $\eta = 0$, and the noise term, $\sqrt{2\lambda\beta^{-1}}\xi_{n+1}$, throughout the exposition. Also, we refer to λ as the *learning rate* and $|\Delta\theta_n^\lambda| := \frac{|G^{(i)}(\theta_n^\lambda, X_{n+1})|}{1+\sqrt{\lambda}|G^{(i)}(\theta_n^\lambda, X_{n+1})|}\left(1+\frac{\sqrt{\lambda}}{\varepsilon+|G^{(i)}(\theta_n^\lambda, X_{n+1})|}\right)$ as the *stepsize* by the convention in Kingma & Ba (2015).

Firstly, the new algorithm utilizes the taming function to control the super-linearly growing gradient. In a region where the loss function is steep and narrow (the gradient is huge), it is ideal for the optimizer to take a small stepsize. This is effectively achieved since the growth of the taming function is proportional to G, but the boosting is close to one when the gradient is huge. The effectiveness of the taming function is confirmed in the motivating example in Section 2. Note that the taming function 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 our new algorithm in solving high-dimensional optimization problems such as the fine tuning of neural network models.

Secondly, we have designed the boosting function to accelerate training speed and prevent the vanishing gradient problem[2]. When the current parameter is located in a region where the loss function is flat (the gradient is small), it is desirable for the optimizer to take a large stepsize. As the gradient gets smaller, the boosting function increases the stepsize of THεO POULA by up to $\sqrt{\lambda}/\varepsilon$, whereas the taming function's contribution decreases. As a result, THεO POULA takes a larger stepsize. In other words, THεO POULA takes a desirable stepsize depending on the magnitude of the gradient. Most importantly, 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 that is evident from our simulation results.

Thirdly, THεO POULA is quickly converted from adaptive learning rate methods to SGD. In the early training phase, THεO POULA certainly behaves like adaptive learning rate methods. Then,

[2] We provide the effectiveness of the boosting function in Appendix E.3 by comparing the performance of THεO POULA with/without the boosting function. The experiment shows that the addition of the boosting function brings a significant improvement in test accuracy across different models and data sets.

when the current position is approaching an optimal solution where $|G^{(i)}|$s are close to zero, the movement of THεO POULA is similar to SGD with a learning rate $(1+\sqrt{\lambda}/\varepsilon)$. Consequently, TH$\varepsilon$O POULA 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 & 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 THεO POULA. Adding properly scaled Gaussian noise allows the new algorithm to escape local minima in a similar manner to the standard SGLD method, see Raginsky et al. (2017).

3.3 Convergence Analysis

We present in this section the main convergence results of THεO POULA to π_β in Wasserstein-1 and Wasserstein-2 distances as defined in (4). The convergence is guaranteed when the learning rate is less than $\lambda_{\max}$, which is given by

$$\lambda_{\max} = \min\left\{\frac{1}{4\eta^2}, \frac{1}{2^{14}\eta^2({}_{8l}\mathcal{C}_{4l})^2}\right\}. \tag{9}$$

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 a η is typically very small ($\eta \ll 1$). Moreover, let $T := 1/\lambda$.

Theorem 3.1 and Corollary 3.1 state the non-asymptotic (upper) bounds between $\mathcal{L}\left(\theta_n^\lambda\right)$ and π_β. An overview of the proofs of our main results can be found in Appendix C.

Theorem 3.1. *Let Assumptions 3.1 and 3.2 hold. Then, for every $0 < \lambda \le \lambda_{\max}$ and $n \in \mathbb{N}$, we have*

$$W_1\left(\mathcal{L}\left(\theta_n^\lambda\right), \pi_\beta\right) \le M_1\sqrt{\lambda} + M_2 e^{-\dot{c}\lambda n},$$

where $\dot{c}$, M_1 and M_2 are constants independent of n and λ. The explicit form of $\dot{c}$, M_1 and M_2 are given in Table 7.

Corollary 3.1. *Let Assumptions 3.1 and 3.2 hold. Then, for every $0 < \lambda \le \lambda_{max}$ and $n \in \mathbb{N}$, we have*

$$W_2\left(\mathcal{L}\left(\theta_n^\lambda\right), \pi_\beta\right) \le M_3\sqrt{\lambda} + z_2\lambda^{\frac{1}{4}} + M_4 e^{-\dot{c}\lambda n},$$

where $\dot{c}$, z_2, M_3 and M_4 are constants independent of n and λ. The explicit form of $\dot{c}$, z_2, M_3 and M_4 are given in Table 7.

We are now concerned with the expected excess risk of THεO POULA generated by (7), so called the optimization error of θ_n^λ, defined as

$$\mathbb{E}[u(\theta_n^\lambda)] - u(\theta^*) \tag{10}$$

where $\theta^* := \arg\min_{\theta\in\mathbb{R}^d} u(\theta)$. To derive the bound of the expected excess risk, it is again decomposed into two parts; $\mathbb{E}[u(\theta_n^\lambda)] - \mathbb{E}[u(\theta_\infty)]$ and $\mathbb{E}[u(\theta_\infty)] - u(\theta^*)$. Here, θ_∞ follows the invariant distribution π_β. The following theorem describes the bound of the expected excess risk of THεO POULA.

Theorem 3.2. *Let Assumptions 3.1 and 3.2 hold. For any $n \in \mathbb{N}$, the expected excess risk of the n-th iterate of THεO POULA (7) is upper bounded by*

$$\mathbb{E}[u(\theta_n^\lambda)] - u(\theta^*) \le M_5 W_2(\mathcal{L}(\theta_n^\lambda, \pi_\beta)) + \frac{1}{\beta}\left[\frac{d}{2}\log\left(\frac{Ke}{A}\left(\frac{B}{d}\beta + 1\right)\right) + \log 2\right]$$

where $W_2(\mathcal{L}(\theta_n^\lambda), \pi_\beta)$ is given in Corollary 3.1 and A, B are given in Remark B.2. and M_5, K are given in Table 7. All the constatns are independent of n and λ.

3.4 Averaged THεO POULA

One notes that Theorem 3.1 implies that THεO POULA converges, under suitable decreasing step size regime, to the invariant measure π_β and thus its performance can be further improved by averaging. It is achieved 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 & Juditsky, 1992). In particular, we use a 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.

Our experiments show that averaged THεO POULA performs better than the other stochastic optimization methods for language modeling tasks. Moreover, while a learning rate decay, which requires additional tuning effort, has to be applied for the other optimizers to obtain their best performance, averaged THεO POULA uses a constant learning rate, which is another practical benefit of our newly proposed algorithm.

4 Empirical performance on real data sets

This section examines the performance of THεO POULA on real data sets by comparing it with those of other stochastic optimization algorithms including Adam (Kingma & Ba (2015)), AdaBelief (Zhuang et al. (2020)), AdamP (Heo et al. (2021)), AdaBound (Luo et al. (2019)), AMSGrad (Reddi et al. (2018)), RMSProp (Tieleman & Hinton (2012)), SWATS (Keskar & Socher (2017)), SGD (with momentum) and ASGD (Merity et al. (2018)). We conduct the following deep learning experiments: image classficiation on CIFAR-10 (Krizhevsky et al.) and CIFAR-100 (Krizhevsk (2009)) and language modeling on Penn Treebank (Marcus et al. (1999)). Each experiment is run three times to compute the mean and standard deviation of the best accuracy on the test dataset. We provide details of the experiments including learning curves and hyperparameter settings in Appendix E.

For our experiments, we consider $\eta = 0$ in (5). This is justified by the fact that some form of dissipativity may exist for specific problems such as the one considered here, although this has not been verified theoretical so far. In Appendix F.2, we perform additional experiments with $\eta \neq 0$, which show a very similar performance by 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 of VGG11 (Simonya & Zisserman (2015)), ResNet34 (Ioffe & Szegedy (2016)) and DenseNet121 (Huang et al. (2017)) on CIFAR-10 and CIFAR-100 in the official implementation of Zhuang et al. (2020). They provide a reliable baseline of the experiments by comparing the performance of various stochastic optimizers with extensive hyperparameter search. We search the optimal hyperparameters for THεO POULA among $\lambda = \{1, 0.5, 0.1, 0.05, 0.01\}$ and $\varepsilon = \{1, 0.1, 0.01\}$. β is chosen among $\{10^8, 10^{10}, 10^{12}\}$ across all the experiments.

Table 2 shows the test accuracy for VGG11, ResNet34 and DenseNet121 on CIFAR-10 and CIFAR-100. As reported in Table 2, our algorithm achieves the highest accuracy and significantly outperforms the other optimizers across all the experiments. In particular, even THεO POULA with the second best hyperparameter is comparable to AdaBelief and outperforms the other methods, validating that the solutions found by THεO POULA yield good generalization performance. Also, the improvement of our algorithm is increasingly prominent as the models and datasets are more complicated and large-scale.

Language modeling We perform language modeling over the Penn Treebank (PTB) with AWD-LSTMs of Merity et al. (2018). It is reported that Non-monotonically Triggered ASGD (NT-ASGD) achieves state-of-the-art performance for the language modeling task with AWD-LSTMs. Motivated by this observation, we consider averaged THεO POULA for the experiment. Due to a limited computation budget, we only test ASGD and AdaBelief rather than investigating all the optimizers in this experiment [3].

For a fair comparison, the averaging scheme has also been applied to AdaBelief, but we have found that it 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 if the model does not attain a new best value for multiple epochs. For ASGD and THεO POULA, a constant learning

[3] Since AdaBelief significantly outperforms the other optimizers including vanilla SGD, AdaBound, Yogi (Zaheer et al. (2018)), Adam, MSVAG (Balles & Hennig (2018)), RAdam, Fromage and AdamW (Loshchilov & Hutter (2019)) in the same experiment, we believe that we do not need to explore all the optimizers.

Table 2: Mean and standard deviation of the best accuracy for VGG11, ResNet34 and DenseNet121 on CIFAR10. THεOPOULA† and THεOPOULA* represent the performances of THεO POULA with the best and second best hyperparameters, respectively.

dataset	CIFAR-10			CIFAR-100		
model	VGG	ResNet	DenseNet	VGG	ResNet	DenseNet
THεO POULA†	**92.31** (0.055)	**95.43** (0.095)	**95.66** (0.066)	**70.31** (0.117)	**77.60** (0.144)	**79.90** (0.133)
THεO POULA*	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)

rate is used without a learning rate decay. In order to compare with the baseline, we apply gradient clipping of 0.25 to all optimizers. See Appendix E for more information.

Table 3 shows that THεO POULA attains the lower test perplexity against the baselines for AWD-LSTM with one, two and three layers. AdaBelief shows a comparable performance with ASGD for 2-layer and 3-layer models.

Our experimental results show that THεO POULA achieves higher accuracy than AdaBelief (known as the state-of-the-art algorithm for many deep learning tasks) on image classification and language modeling tasks for various deep learning models. Furthermore, it is easier to tune parameters of THεO POULA since the number of hyperparmeters for THεO POULA is less than that of Adam-type optimizers.

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

# of layers	1-layer	2-layer	3-layer
THεO POULA	**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)

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 THεO POULA is provided along with full theoretical guarantees for obtaining the best known convergence rates. Our experiments confirm that THεO POULA outperforms other popular stochastic optimization methods.

We believe that 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.

References

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.

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*, pp. 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.J. Oh, D. Han 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*, pp. 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*, pp. 1177–1182, 1980.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference on Machine Learning*, pp. 448–456, 2015.

S. Ioffe and C. Szegedy. Deep residual learning for image recognition. *IEEE conference on computer vision and pattern recognition*, pp. 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.

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.

M.P. Marcus, B. Santorini, M.A. Marcinkiewicz, and Ann 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.

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*, pp. 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.

Appendix

A Details of the Experiment in Section 2

This section provides the necessary theoretical details of the experiment in Section 2. We continue to consider the optimization problem (2). One calculates that

$$u(\theta) = \begin{cases} \theta^{30} + \frac{7}{4}\theta^2, & |\theta| \leq 1, \\ \theta^{30} + \frac{7}{4}(2|\theta| - 1), & |\theta| > 1 \end{cases}$$

and

$$u'(\theta) = \begin{cases} 30\theta^{29} + \frac{7}{2}\theta, & |\theta| \leq 1, \\ 30\theta^{29} + \frac{7}{2}sgn(\theta), & |\theta| > 1. \end{cases}$$

Note that $u(\theta)$ and $u'(\theta)$ are continuous since $u(1) = \lim_{\theta\downarrow 1} u(\theta) = \frac{11}{4}$, $u(-1) = \lim_{\theta\uparrow -1} = \frac{11}{4}$, $u'(1) = \lim_{\theta\downarrow -1} u'(1) = \frac{67}{2}$ and $u'(-1) = \lim_{\theta\uparrow 1} u'(-1) = -\frac{67}{2}$. Therefore, the minimum value is attained at $\theta = 0$.

To show that G is locally Lipschitz continuous, we check that for $|\theta|, |\theta'| > 1$ and $x \in \mathbb{R}^d$,

$$\begin{aligned} |G(\theta, x) - G(\theta', x)| &\leq (2 + 2\mathbb{1}_{x\leq 1})|sgn(\theta) - sgn(\theta')| + 30|\theta^{29} - \theta'^{29}| \\ &\leq 34(1 + |\theta| + |\theta'|)^{28}|\theta - \theta'|. \end{aligned}$$

For $|\theta|, |\theta| \leq 1$, we have

$$\begin{aligned} |G(\theta, x) - G(\theta', x)| &\leq (2 + 2\mathbb{1}_{x\leq 1})|\theta - \theta'| + 30|\theta^{29} - \theta'^{29}| \\ &\leq 34(1 + |\theta| + |\theta'|)^{28}|\theta - \theta'|. \end{aligned}$$

For $|\theta| \leq 1$, $|\theta| > 1$, we obtain

$$\begin{aligned} |G(\theta, x) - G(\theta', x)| &\leq (2 + 2\mathbb{1}_{x\leq 1})|\theta - sgn(\theta')| + 30|\theta^{29} - \theta'^{29}| \\ &\leq (2 + 2\mathbb{1}_{x\leq 1})|\theta - \theta'| + 30|\theta^{29} - \theta'^{29}| \\ &\leq 34(1 + |\theta| + |\theta'|)^{28}|\theta - \theta'| \end{aligned}$$

where the second inequality follows from the following relations

$$\begin{aligned} \theta - \theta' \leq \theta - 1 \leq 0, &\quad \text{for } \theta' > 1, \\ 0 \leq \theta + 1 \leq \theta - \theta', &\quad \text{for } \theta' < -1. \end{aligned}$$

B Key Observations from Assumption 3.1 and 3.2

This section introduces some useful general results, that can be obtained from Assumption 3.1 and 3.2. Note that some of the below observations can be also found in Zhang et al. (2019) and Lovas et al. (2020). However, to make our paper self-contained, we record all the results which are necessary for the convergence analysis.

Remark B.1. *From Assumption 3.1, one observes that for all $\theta \in \mathbb{R}^d$ and $x \in \mathbb{R}^m$*

$$|G(\theta, x)| \leq K(x)(1 + |\theta|^q),$$

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

Remark B.2. *From Assumptions 3.1 and 3.2, one obtains that*

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

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

$$\langle \theta,\, h(\theta)\rangle \geq A|\theta|^2 - B. \tag{B.1}$$

Proposition B.1. *(Lovas et al. (2020)) Let Assumptions 3.1 and 3.2 hold. Then, for every θ, $\theta' \in \mathbb{R}^d$,*

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

where $a = L_1\mathbb{E}[(1 + |X_0|)^\rho](1 + 2|R|)^{q-1}$ and R is given by

$$R = \max\left\{\left(\frac{2^{3(q-1)+1}L_1\mathbb{E}[(1 + |X_0|)^\rho]}{\eta}\right)^{\frac{1}{2r-1}}, \left(\frac{2^q L_1\mathbb{E}[(1 + |X_0|)^\rho]}{\eta}\right)^{\frac{1}{2r}}\right\}.$$

Proposition B.2. *(Lovas et al. (2020)) Let Assumptions 3.1 and 3.2 hold. Then, one obtains that*

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

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

Remark B.3. *From Assumption 3.1 and the definition of H and $H_{\lambda,c}$, one obtains that for $\theta\in\mathbb{R}^d$, $x\in\mathbb{R}^m$ and $i = 1, 2, \cdots, d$,*

$$\begin{aligned}
|H^{(i)}(\theta,x)-H^{(i)}_{\lambda,c}(\theta,x)| &\leq \left|G^{(i)}(\theta,x)-\frac{G^{(i)}(\theta,x)}{1+\sqrt{\lambda}|G^{(i)}(\theta,x)|}\left(1+\frac{\sqrt{\lambda}}{\varepsilon+|G^{(i)}(\theta,x)|}\right)\right| \\
&+ \left|\eta\theta^{(i)}|\theta|^{2r}-\eta\frac{\theta^{(i)}|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}}\right| \\
&\leq |G^{(i)}(\theta,x)|\frac{\sqrt{\lambda}|G^{(i)}(\theta,x)|}{1+\sqrt{\lambda}|G^{(i)}(\theta,x)|}+\frac{\sqrt{\lambda}|G^{(i)}(\theta,x)|}{(1+\sqrt{\lambda}|G^{(i)}(\theta,x)|)(\varepsilon+|G^{(i)}(\theta,x)|)} \\
&+ \eta|\theta^{(i)}||\theta|^{2r}\left|\frac{\sqrt{\lambda}|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}}\right| \\
&\leq \sqrt{\lambda}|G^{(i)}(\theta,x)|^2+\sqrt{\lambda}+\sqrt{\lambda}\eta|\theta^{(i)}||\theta|^{4r}
\end{aligned}$$

which implies that

$$\begin{aligned}
|H(\theta,x)-H_{\lambda,c}(\theta,x)|^2 &= \sum_{i=1}^{d}\left[\sqrt{\lambda}|G^{(i)}(\theta,x)|^2+\sqrt{\lambda}+\sqrt{\lambda}\eta|\theta^{(i)}||\theta|^{4r}\right]^2 \\
&\leq 3\lambda\sum_{i=1}^{d}\left[|G^{(i)}(\theta,x)|^4+1+\eta^2|\theta^{(i)}|^2|\theta|^{8r}\right] \\
&\leq 3\lambda\left[\left(\sum_{i=1}^{d}|G^{(i)}(\theta,x)|^2\right)^2+d+\eta^2|\theta|^{8r+2}\right] \\
&\leq 3\lambda\left[|G(\theta,x)|^4+d+\eta^2|\theta|^{8r+2}\right] \\
&\leq 3\lambda\left[8|K(x)|^4(1+|\theta|^{4q})+d+\eta^2|\theta|^{8r+2}\right].
\end{aligned}$$

Remark B.4. *From Assumption 3.1 and the definition of H and $H_{\lambda,c}$, one calculates that for all $\theta\in\mathbb{R}^d$ and $x\in\mathbb{R}^m$,*

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

and

$$\begin{aligned}
|H_{\lambda,c}(\theta,x)|^2 &= \sum_{i=1}^{d}\left[\frac{G^{(i)}(\theta,x)}{1+\sqrt{\lambda}|G^{(i)}(\theta,x)|}\left(1+\frac{\sqrt{\lambda}}{\varepsilon+|G^{(i)}(\theta,x)|}\right)+\eta\frac{\theta^{(i)}|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}}\right]^2 \\
&\leq \sum_{i=1}^{d}\left[\frac{|G^{(i)}(\theta,x)|}{1+\sqrt{\lambda}|G^{(i)}(\theta,x)|}+\frac{\sqrt{\lambda}|G^{(i)}(\theta,x)|}{(1+\sqrt{\lambda}|G^{(i)}(\theta,x)|)(\varepsilon+|G^{(i)}(\theta,x)|)}+\eta\frac{|\theta^{(i)}||\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}}\right]^2 \\
&\leq \sum_{i=1}^{d}\left(|G^{(i)}(\theta,x)|+\sqrt{\lambda}+\eta|\theta^{(i)}||\theta|^{2r}\right)^2 \\
&\leq 3\sum_{i=1}^{d}\left(|G^{(i)}(\theta,x)|^2+\lambda+\eta^2|\theta^{(i)}|^2|\theta|^{4r}\right) \\
&\leq 3|G(\theta,x)|^2+3\lambda d+3\eta^2|\theta|^{4r+2} \\
&\leq 6|K(x)|^2(1+|\theta|^{2q})+3\lambda d+3\eta^2|\theta|^{4r+2}.
\end{aligned}$$

C Overview of the Proofs

This section provides an overview of the proofs of our main results. We begin by introducing suitable Lyapunov functions and auxiliary processes to analyze the convergence of our newly introduced algorithm. For each $m \geq 1$, define the Lyapunov function V_m by

$$V_m(\theta) := (1+|\theta|^2)^{\frac{m}{2}}, \quad \theta \in \mathbb{R}^d \tag{C.1}$$

and similarly $v_m(x) = (1+x^2)^{\frac{m}{2}}$ for any real $x \geq 0$. Both functions are continuously differentiable and $\lim_{|\theta|\to\infty} \nabla V_m(\theta)/V_m(\theta) = 0$. Also, define $Z_t^\lambda = Z_{\lambda t}$, which is the time-changed Lagevin dynamics governed by

$$dZ_t^\lambda = -\lambda h(Z_t^\lambda)dt + \sqrt{2\beta^{-1}\lambda} d\tilde{B}_t^\lambda \tag{C.2}$$

where $\tilde{B}_t^\lambda = B_{\lambda t}/\sqrt{\lambda}$ is a Brownian motion.

We next define the continuous-time interpolation of the new algorithm, see (7), as

$$\mathrm{d}\bar{\theta}_t^\lambda = -\lambda H_\lambda\left(\bar{\theta}_{\lfloor t\rfloor}^\lambda, X_{\lceil t\rceil}\right)\mathrm{d}t + \sqrt{2\lambda\beta^{-1}}\mathrm{d}\tilde{B}_t^\lambda \tag{C.3}$$

with initial condition $\bar{\theta}_0^\lambda = \theta_0^\lambda$. Henceforth, $\lfloor x\rfloor$ denotes the integer part of a positive real x and $\lceil x\rceil = \lfloor x\rfloor + 1$.

Remark C.1. *Due to the homogeneous nature of the coefficients of the continuous-time interpolation of THεO POULA (C.3) and when one selects a version of the driving Brownian motion such that it coincides with ξ_n at grid points, it follows that the interpolated process (C.3) equals the process of THεO POULA (7) almost surely at grid points, i.e. $\bar{\theta}_n^\lambda = \theta_n^\lambda$ (a.s), $\forall n \in \mathbb{N}$.*

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

$$\mathrm{d}\zeta_t^{s,v,\lambda} = -\lambda h\left(\zeta_t^{s,v,\lambda}\right)\mathrm{d}t + \sqrt{2\lambda\beta^{-1}}\mathrm{d}\tilde{B}_t^\lambda \tag{C.4}$$

with initial condition $\zeta_s^{s,v,\lambda} := v, v \in \mathbb{R}^d$. Let us also define $T := \frac{1}{\lambda}$, which allows us to create suitable subintervals on the positive real line in order to compare the behaviour of the aforementioned processes at each such interval.

Definition C.1. *Fix $k \in \mathbb{N}$ and define $\bar{\zeta}_t^{\lambda,k} := \zeta_t^{kT,\bar{\theta}_{kT}^\lambda,\lambda}$ where $\zeta_t^{kT,\bar{\theta}_{kT}^\lambda,\lambda}$ is defined in (C.4).*

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

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

for $j = 1, 2$.

C.1 Primary estimates

We collect first the necessary estimates in order to obtain (upper) bounds for $W_1\left(\mathcal{L}\left(\theta_t^\lambda\right), \pi_\beta\right)$ and $W_2\left(\mathcal{L}\left(\theta_t^\lambda\right), \pi_\beta\right)$. All proofs of the lemmas in this section can be found in Appendix D. The following two lemmas provide, uniform in n, moment estimates of the process $(\theta_n^\lambda)_{n\geq 1}$.

Lemma C.1. *Let Assumptions 3.1 and 3.2 hold. Then, there exist $M_0 > 0$ and $\lambda_{\max}$, which is defined in (9), such that for any $\lambda \in (0, \lambda_{\max})$ and any $n \in \mathbb{N}$,*

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

and, moreover,

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

Lemma C.2. *Let Assumptions 3.1 and 3.2 hold. Then, there exist $M_0 > 0$ and $\lambda_{\max}$, which is defined in (9), such that for any $\lambda \in (0,\ \lambda_{\max})$, $n \in \mathbb{N}$, and $p \in [1, 8(2r+1)]$,*

$$\mathbb{E}|\theta^{\lambda}_{n+1}|^{2p} \leq (1-\eta^2\lambda)^n \mathbb{E}|\theta^{\lambda}_0|^{2p} + \frac{A_p}{\eta^2}$$

and

$$\sup_n \mathbb{E}|\theta^{\lambda}_{n+1}|^{2p} \leq \mathbb{E}|\theta^{\lambda}_0|^{2p} + \frac{A_p}{\eta^2}$$

where A_p is given in Table 7.

Lemma C.3. *Let Assumptions 3.1 and 3.2 hold. Then, there exist $M_0 > 0$ and $\lambda_{\max}$, which is defined in (9), such that for any $\lambda \in (0,\ \lambda_{\max})$ and $n \in \mathbb{N}$,*

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

where A_2, i.e. A_p for $p = 2$, is given in Table 7.

Proof. From the definition of the Lyapunov function and Remark C.1, we have

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

□

Moreover, the necassary moment bounds hold also for the auxiliary process $\{\bar{\zeta}^{\lambda,n}_t\}_{t \geq nT}$.

Lemma C.4. *(Lemma 3.5. of Chau et al. (2019)) Let Assumptions 3.1 and 3.2 hold. Then,*

$$\begin{aligned}\mathbb{E}[V_2(\bar{\zeta}^{\lambda,n}_t)] &\leq \mathbb{E}[V_2(\theta_0)] + \frac{\tilde{c}(2)}{\bar{c}(2)} + 2(C_X\eta^{-1} + 2M_0^2(2+\eta) + 2d(\eta\beta)^{-1}\sqrt{\lambda_{\max}}) + 1, \\ \mathbb{E}[V_4(\bar{\zeta}^{\lambda,n}_t)] &\leq 2\mathbb{E}|\theta_0|^4 + 2 + 2\frac{A_2}{\eta^2} + \frac{\tilde{c}(4)}{\bar{c}(4)},\end{aligned}$$

where $\bar{c}(p)$, $\tilde{c}(p)$ are given in Table 7.

Let $\mathcal{P}_{V_2}$ denote the subset of $\mathcal{P}(\mathbb{R}^d)$ such that every $\mu \in \mathcal{P}_{V_2}$ satisfies $\int_{\mathbb{R}^d} V_2(\theta)\mu(d\theta) < \infty$. Moreover, let the following functional be considered

$$w_{1,2}(\mu,\nu) := \inf_{\zeta \in \mathcal{C}(\mu,\nu)} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} [1 \wedge |\theta - \theta|'] \left[(1 + V_2(\theta) + V_2(\theta')\right) \zeta(\mathrm{d}\theta \mathrm{d}\theta') \tag{C.5}$$

where $\mathcal{C}(\mu,\nu)$ is defined in (4). The following lemma states the contraction property of the Langevin SDE (C.2) in $w_{1,2}$, which yields the desired result for $W_1(\mathcal{L}(Z^{\lambda}_n), \pi_\beta)$.

Lemma C.5. *(Proposition 3.14 of Chau et al. (2019)) Let $Z'_t, t \in \mathbb{R}_+$ be the solution of the Langevin SDE (6) with initial condition $Z'_0 = \theta'_0$ which is independent of $\mathcal{G}_\infty$ and $|\theta'_0| \in L^2$. Then,*

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

where $w_{1,2}$ is defined in (C.5).

The following two Lemmas combined establish the required $W_1(\mathcal{L}(\bar{\theta}^{\lambda}_t), \mathcal{L}(Z^{\lambda}_t))$ estimate. One recalls first that for any $t > 0$, there exists a unique integer m such that $t \in [mT, (m+1)T)$.

Lemma C.6. *Let Assumptions 3.1 and 3.2 hold. Then, for $0 < \lambda < \lambda_{max}$, one obtains*

$$W_2\left(\mathcal{L}\left(\bar{\theta}^{\lambda}_t\right), \mathcal{L}\left(\bar{\zeta}^{\lambda,m}_t\right)\right) \leq \sqrt{\lambda}\sqrt{e^{3a}(C_1 + C_2 + C_3)}$$

where C_1, C_2, C_3 are given explicitly in Table 7.

Lemma C.7. *Let Assumptions 3.1 and 3.2 hold. Then, for $0 < \lambda < \lambda_{max}$, one obtains*

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

where z_1 is given explicitly in Table 7.

Lemma C.8. *Let Assumptions 3.1 and 3.2 hold. Then, for $0 < \lambda < \lambda_{max}$, one obtains*

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

where z_2 is given explicitly in Table 7.

C.2 Proofs of main results

It is assumed throughout the paper that the random variable θ_0, $\mathcal{G}_\infty := \sigma\left(\cup_{n\in\mathbb{N}}\mathcal{G}_n\right)$ and $(\xi_n)_{n\in\mathbb{N}}$ are independent.

Proof of Theorem 3.1. Observe that $W_1\left(\mathcal{L}\left(\theta_n^\lambda\right), \mathcal{L}\left(Z_t^\lambda\right)\right)$ is decomposed as follows:

$$W_1\left(\mathcal{L}\left(\theta_n^\lambda\right), \pi_\beta\right) \le W_1\left(\mathcal{L}\left(\bar\theta_n^\lambda\right), \mathcal{L}\left(Z_n^\lambda\right)\right) + W_1\left(\mathcal{L}\left(Z_n^\lambda\right), \pi_\beta\right).$$

Note that there exists a unique integer m such that $n \in [mT, (m+1)T)$. Thus, from the results of Lemma C.6 and C.7, the first term in the right-hand side is estimated

$$\begin{aligned}
W_1\left(\mathcal{L}\left(\bar\theta_n^\lambda\right), \mathcal{L}\left(Z_n^\lambda\right)\right) &\le W_1\left(\mathcal{L}\left(\bar\theta_n^\lambda\right), \mathcal{L}\left(\bar\zeta_n^{\lambda,m}\right)\right) + W_1\left(\mathcal{L}\left(\bar\zeta_n^{\lambda,m}\right), \mathcal{L}\left(Z_n^\lambda\right)\right)\\
&\le W_2\left(\mathcal{L}\left(\bar\theta_n^\lambda\right), \mathcal{L}\left(\bar\zeta_n^{\lambda,m}\right)\right) + W_1\left(\mathcal{L}\left(\bar\zeta_n^{\lambda,m}\right), \mathcal{L}\left(Z_n^\lambda\right)\right)\\
&\le \sqrt{\lambda}\left(\sqrt{e^{3a}(C_1+C_2+C_3)} + z_1\right).
\end{aligned}$$

Consequently, we derive

$$\begin{aligned}
W_1\left(\mathcal{L}\left(\theta_n^\lambda\right), \pi_\beta\right) &\le \sqrt{\lambda}\left(\sqrt{e^{3a}(C_1+C_2+C_3)} + z_1\right) + w_{1,2}\left(\mathcal{L}\left(Z_n^\lambda\right), \pi_\beta\right)\\
&\le \sqrt{\lambda}\left(\sqrt{e^{3a}(C_1+C_2+C_3)} + z_1\right) + \hat{c}e^{-\dot{c}\lambda n} w_{1,2}(\theta_0, \pi_\beta)\\
&\le \sqrt{\lambda}\left(\sqrt{e^{3a}(C_1+C_2+C_3)} + z_1\right) + \hat{c}e^{-m\dot{c}}\left[1 + \mathbb{E}[V_2(\theta_0)] + \int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(d\theta)\right]
\end{aligned}$$

where Remark C.5 is used for the first inequality. □

Proof of Corollary 3.1. Let $n \in [mT, (m+1)T)$. Then, Lemma C.6 and C.8 and Remark C.5 yield that

$$\begin{aligned}
W_2\left(\mathcal{L}\left(\theta_n^\lambda\right), \pi_\beta\right) &\le W_2\left(\mathcal{L}\left(\bar\theta_n^\lambda\right), \mathcal{L}\left(Z_n^\lambda\right)\right) + W_2\left(\mathcal{L}\left(Z_n^\lambda\right), \pi_\beta\right)\\
&\le W_2\left(\mathcal{L}\left(\bar\theta_n^\lambda\right), \mathcal{L}\left(\bar\zeta_n^{\lambda,m}\right)\right) + W_2\left(\mathcal{L}\left(\bar\zeta_n^{\lambda,m}\right), \mathcal{L}\left(Z_n^\lambda\right)\right) + W_2\left(\mathcal{L}\left(Z_n^\lambda\right), \pi_\beta\right)\\
&\le \sqrt{e^{3a}(C_1+C_2+C_3)}\sqrt{\lambda} + z_2\lambda^{\frac{1}{4}} + \sqrt{2w_{1,2}(\mathcal{L}(Z_t^\lambda), \pi_\beta)}\\
&\le \sqrt{e^{3a}(C_1+C_2+C_3)}\sqrt{\lambda} + z_2\lambda^{\frac{1}{4}} + \sqrt{\hat{c}}e^{-\dot{c}\lambda n/2}\sqrt{2w_{1,2}(\theta_0, \pi_\beta)}\\
&\le \sqrt{e^{3a}(C_1+C_2+C_3)}\sqrt{\lambda} + z_2\lambda^{\frac{1}{4}} + \sqrt{2\hat{c}e^{-\dot{c}m/2}\left(1 + \mathbb{E}\left[V_2\left(\theta_0\right)\right] + \int_{\mathbb{R}^d} V_2(\theta)\pi_\beta(d\theta)\right)}.
\end{aligned}$$

□

Proof of Theorem 3.2. We begin by decomposing expected excess risk (10) as follows:

$$\mathbb{E}[u(\theta_n^\lambda)] - u(\theta^*) \le \mathbb{E}[u(\theta_n^\lambda)] - \mathbb{E}[u(\theta_\infty)] + \mathbb{E}[u(\theta_\infty)] - u(\theta^*).$$

Let us focus on estimating the first part, $\mathbb{E}[u(\theta_n^\lambda)] - \mathbb{E}[u(\theta_\infty)]$. Observe that for $\theta \in \mathbb{R}^d$

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

by separating the cases $|\theta| \le 1$ and $|\theta| > 1$ where $r_1 = \mathbb{E}[K(X_0)] + \eta$ due to Remark B.1. Then, we have

$$\begin{aligned}
u(w) - u(v) &= \int_0^1 \langle \nabla u((1-t)v + tw),\, w - v\rangle dt\\
&\le \int_0^1 |\nabla u((1-t)v + tw)||w - v| dt\\
&\le \int_0^1 \left(a_1(1-t)^l|v|^l + a_1 t^l|w|^l + 2\mathbb{E}[K(X_0)]\right)|w - v| dt\\
&= \left(\frac{a_1}{l+1}|v|^l + \frac{a_1}{l+1}|w|^l + 2\mathbb{E}[K(X_0)]\right)|w - v| \qquad \text{(C.6)}
\end{aligned}$$

where $l = 2r + 1$ and $a_1 = 2^l r_1 = 2^l(\mathbb{E}|K(X_0)| + \eta)$. Let P denote the coupling between μ and ν that achieves $W_2(\mu, \nu)$ with $\mu = \mathcal{L}(\theta_n^\lambda)$ and $\nu = \mathcal{L}(\theta_\infty)$. Then, from (C.6), we obtain

$$\begin{aligned}
\mathbb{E}u(\theta_n^\lambda) - \mathbb{E}u(\theta_\infty) &= \mathbb{E}_P[u(\theta_n^\lambda) - u(\theta_\infty)] \\
&\leq \mathbb{E}_P\left[\left(\frac{a_1}{l+1}|\theta_n^\lambda|^l + \frac{a_1}{l+1}|\theta_\infty|^l + 2\mathbb{E}_P[K(X_0)]\right)|\theta_n^\lambda - \theta_\infty|\right] \\
&\leq \sqrt{\mathbb{E}_P\left[\left(\frac{a_1}{l+1}|\theta_n^\lambda|^l + \frac{a_1}{l+1}|\theta_\infty|^l + 2\mathbb{E}[K(X_0)]\right)^2\right]}\sqrt{\mathbb{E}_P|\theta_n^\lambda - \theta_\infty|^2} \\
&\leq \left(\frac{a_1}{l+1}\sqrt{\mathbb{E}|\theta_n^\lambda|^{2l}} + \frac{a_1}{l+1}\sqrt{\mathbb{E}|\theta_\infty|^{2l}} + 2\mathbb{E}[K(X_0)]\right)W_2\left(\mathcal{L}(\theta_n^\lambda), \pi_\beta\right) \\
&\leq \left(\frac{a_1}{l+1}\sqrt{\mathbb{E}|\theta_0|^{2l} + \frac{A_l}{\eta^2}} + \frac{a_1}{l+1}\sqrt{\mathbb{E}|\theta_\infty|^{2l}} + 2\mathbb{E}[K(X_0)]\right) \\
&\quad \times W_2\left(\mathcal{L}(\theta_n^\lambda), \pi_\beta\right)
\end{aligned} \tag{C.7}$$

where we have used Lemma C.2 for the last inequality.

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

$$\begin{aligned}
\mathbb{E}u(\theta_\infty) - u(\theta^*) &\leq \frac{1}{\beta}\left(-\int_{\mathbb{R}^d}\frac{e^{-\beta u(\theta)}}{\Lambda}\log\frac{e^{-\beta u(\theta)}}{\Lambda}d\theta - \log\Lambda\right) - u^* \\
&\leq \frac{d}{2\beta}\log\left(\frac{2\pi e(B + d/\beta)}{Ad}\right) - \frac{\log\Lambda}{\beta} - u^*
\end{aligned} \tag{C.8}$$

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

Using (B.1), 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 Proposition B.2, we further obtain

$$\begin{aligned}
-\beta(u(\theta^*) - u(w)) &= \beta|u(\theta^*) - u(w)| \\
&\leq \beta\int_0^1 \frac{1}{t-1}|\langle h(w + t(\theta^* - w)) - h(\theta^*),\, w - \theta^* + t(\theta^* - w)\rangle|dt \\
&\leq \beta L\mathbb{E}(1 + |X_0|)^\rho\int_0^1 (1 + |w + t(\theta^* - w)| + |\theta^*|)^l(1-t)|w - \theta^*|^2 dt \\
&\leq \beta L\mathbb{E}(1 + |X_0|)^\rho\int_0^1 (1 + |w| + |\theta^* - w| + |\theta^*|)^l(1-t)|w - \theta^*|^2 dt \\
&= \beta L\mathbb{E}(1 + |X_0|)^\rho(1 + 2|\theta^*| + 2|\theta^* - w|)^l\frac{|w - \theta^*|^2}{2}
\end{aligned}$$

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

Define $R_0 := \max\{\sqrt{B/A}, \sqrt{2d/(\beta L\mathbb{E}(1+|X_0|)^\rho)}\}$ and $\overline{\mathbf{B}}_r(p) = \{x \in \mathbb{R}^d||x-p| > r\}$. Then, from the above inequality, one further calculates

$$
\begin{aligned}
\frac{\log \Lambda}{\beta} &= -u(\theta^*) + \frac{1}{\beta}\log\int_{\mathbb{R}^d} e^{\beta(u(\theta^*)-u(w))}dw \\
&\geq -u(\theta^*) + \frac{1}{\beta}\log\int_{\mathbb{R}^d} e^{-\beta L\mathbb{E}(1+|X_0|)^\rho(1+2|\theta^*|+2|\theta^*-w|)^l\frac{|w-\theta^*|^2}{2}}dw \\
&\geq -u(\theta^*) + \frac{1}{\beta}\log\int_{\overline{\mathbf{B}}_{R_0}(\theta^*)} e^{-\beta L\mathbb{E}(1+|X_0|)^\rho(1+4R_0)^l\frac{|w-\theta^*|^2}{2}}dw \\
&= -u(\theta^*) + \frac{1}{\beta}\log\left[\left(\frac{2\pi}{\beta K}\right)^{d/2}\int_{\overline{\mathbf{B}}_{R_0}(\theta^*)} f_X(w)dw\right] \\
&\geq -u(\theta^*) + \frac{1}{\beta}\log\left(\frac{1}{2}\left(\frac{2\pi}{K\beta}\right)^{d/2}\right) \qquad \text{(C.9)}
\end{aligned}
$$

where $K = L\mathbb{E}(1+|X_0|)^\rho(1+4R_0)^l$ 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:

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

Combining (C.8) and (C.9), we derive

$$
\begin{aligned}
\mathbb{E}u(\theta_\infty) - u(\theta^*) &\leq \frac{d}{2\beta}\log\left(\frac{2\pi e(B+d/\beta)}{Ad}\right) - \frac{1}{\beta}\log\left(\frac{1}{2}\left(\frac{2\pi}{K\beta}\right)^{d/2}\right) \\
&\leq \frac{1}{\beta}\left[\frac{d}{2}\log\left(\frac{Ke}{A}\left(\frac{B}{d}\beta+1\right)\right)+\log 2\right]. \qquad \text{(C.10)}
\end{aligned}
$$

Consequently, from (C.7) and (C.10), we derive

$$
\begin{aligned}
\mathbb{E}u(\theta_n^\lambda) - u(\theta^*) &\leq M_5 W_2(\mathcal{L}(\theta_n^\lambda, \pi_\beta)) \\
&+ \frac{1}{\beta}\left[\frac{d}{2}\log\left(\frac{Ke}{A}\left(\frac{B}{d}\beta+1\right)\right)+\log 2\right]
\end{aligned}
$$

where $M_5 = \frac{a_1}{l+1}\sqrt{\mathbb{E}|\theta_0|^{2l}+\frac{A_l}{\eta^2}} + \frac{a_1}{l+1}\sqrt{\mathbb{E}|\theta_\infty|^{2l}} + 2\mathbb{E}[K(X_0)]$. □

D Proofs of Lemmas in Appendix C

Proof of Lemma C.1. Define $\widehat{G}^{(i)}_{\lambda,c}(\theta,x)=\frac{G^{(i)}(\theta,x)}{1+\sqrt{\lambda}|G^{(i)}(\theta,x)|}\left(1+\frac{\sqrt{\lambda}}{\varepsilon+|G^{(i)}(\theta,x)|}\right)$, which is part of the adaptive gradient of $H^{(i)}_{\lambda,c}(\theta,x)$. One observes that $i\in\{1,\cdots,d\}$

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

to obtain

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

for all $x\in\mathbb{R}^m$ and $\theta\in\mathbb{R}^d$. Then,

$$2\lambda\mathbb{E}\Big[\langle\frac{\theta^\lambda_n}{|\theta^\lambda_n|^2},\,H_{\lambda,c}(\theta^\lambda_n,X_{n+1})\rangle\Big|\theta^\lambda_n\Big] \ge -2\Big(\sqrt{\lambda}+\lambda^{\frac{3}{2}}\Big)\frac{\sqrt{d}}{|\theta^\lambda_n|}+2\eta\lambda\frac{|\theta^\lambda_n|^{2r}}{1+\sqrt{\lambda}|\theta^\lambda_n|^{2r}}. \quad (D.2)$$

On the other hand, due to (D.1),

$$\begin{aligned}|H_{\lambda,c}(\theta,x)|^2=\langle H_{\lambda,c}(\theta,x),\,H_{\lambda,c}(\theta,x)\rangle &= \sum_{i=1}^{d}\Big(\widehat{G}^{(i)}_{\lambda,c}(\theta,x)+\eta\frac{\theta^{(i)}|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}}\Big)^2\\ &\le \sum_{i=1}^{d}\Big(2|\widehat{G}^{(i)}_{\lambda,c}(\theta,x)|^2+2\eta^2\frac{|\theta^{(i)}|^2|\theta|^{4r}}{(1+\sqrt{\lambda}|\theta|^{2r})^2}\Big)\\ &\le 2d\Big(\frac{1}{\sqrt{\lambda}}+\sqrt{\lambda}\Big)^2+2\eta^2\frac{|\theta|^{4r+2}}{(1+\sqrt{\lambda}|\theta|^{2r})^2}\\ &\le 4d\Big(\frac{1}{\lambda}+\lambda\Big)+2\eta^2|\theta|^2\frac{|\theta|^{4r}}{(1+\sqrt{\lambda}|\theta|^{2r})^2}. \quad (D.3)\end{aligned}$$

which yields that

$$\begin{aligned}2\lambda\mathbb{E}\Big[-\frac{\lambda}{2|\theta^\lambda_n|^2}|H_{\lambda,c}(\theta^\lambda_n,X_{n+1})|^2\Big|\theta^\lambda_n\Big] &\ge -2\lambda\Big(2d\frac{(1+\lambda^2)}{|\theta^\lambda_n|^2}+\eta^2\frac{\lambda|\theta|^{4r}}{(1+\sqrt{\lambda}|\theta|^{2r})^2}\Big)\\ &\ge -4\lambda d\frac{(1+\lambda^2)}{|\theta^\lambda_n|^2}-2\lambda\eta^2. \quad (D.4)\end{aligned}$$

From (D.2) and (D.4), one calculates that

$$\begin{aligned}&2\lambda\mathbb{E}\Big[\langle\frac{\theta^\lambda_n}{|\theta^\lambda_n|^2},\,H_{\lambda,c}(\theta^\lambda_n,X_{n+1})\rangle-\frac{\lambda}{2|\theta^\lambda_n|^2}|H_{\lambda,c}(\theta^\lambda_n,X_{n+1})|^2\Big|\theta^\lambda_n\Big]\\ \ge\;& -2\Big(\sqrt{\lambda}+\lambda^{\frac{3}{2}}\Big)\frac{\sqrt{d}}{|\theta^\lambda_n|}+2\eta\lambda\frac{|\theta^\lambda_n|^{2r}}{1+\sqrt{\lambda}|\theta^\lambda_n|^{2r}}-4\lambda d\frac{(1+\lambda^2)}{|\theta^\lambda_n|^2}-2\lambda\eta^2=:f(\theta^\lambda_n).\end{aligned}$$

Since $f(\theta)$ tends to $2\eta\sqrt{\lambda} - 2\lambda\eta^2$ as $|\theta| \to \infty$, there exists $M_0 > 0$ such that

$$f(\theta_n^\lambda) \geq \eta\sqrt{\lambda} - \lambda\eta^2 = \eta\sqrt{\lambda}(1 - \sqrt{\lambda}\eta)$$

for all $|\theta_n^\lambda| \geq M_0$ and $\lambda < \frac{1}{\eta^2}$. Moreover, for $\lambda \leq \frac{1}{4\eta^2}$, it can rewritten as there exists $M_0 > 0$ such that

$$f(\theta_n^\lambda) \geq \eta\sqrt{\lambda} - \lambda\eta^2 = \frac{\eta\sqrt{\lambda}}{2} \tag{D.5}$$

for all $|\theta_n^\lambda| \geq M_0$.

Therefore,

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

implying that

$$\begin{aligned}
&\mathbb{E}\left[|\theta_{n+1}^\lambda|^2\mathbf{1}_{|\theta_n^\lambda|\geq M_0}\middle|\theta_n^\lambda\right] \\
= \;& \mathbb{E}\left[\left(|\theta_n^\lambda|^2 - 2\lambda\langle\theta_n^\lambda,\, H_{\lambda,c}(\theta_n, X_{n+1})\rangle + \lambda^2|H_{\lambda,c}(\theta_n, X_{n+1})|^2 + \frac{2\lambda}{\beta}|\xi_{n+1}|^2\right)\mathbf{1}_{|\theta_n^\lambda|\geq M_0}\middle|\theta_n^\lambda\right] \\
\leq \;& \left(1 - \frac{\eta\sqrt{\lambda}}{2}\right)|\theta_n^\lambda|^2 + \frac{2\lambda d}{\beta}.
\end{aligned} \tag{D.6}$$

Let us consider the case of $|\theta_n^\lambda| < M_0$. From the fact that

$$\begin{aligned}
\langle\theta,\, H_{\lambda,c}(\theta, x)\rangle &= \sum_{i=1}^d \theta^{(i)} \cdot \widehat{G}_{\lambda,c}^{(i)}(\theta, x) + \eta\frac{|\theta|^{2r+2}}{1 + \sqrt{\lambda}|\theta|^{2r}} \\
&\leq \sum_{i=1}^d |\theta^{(i)}|\left(\frac{1}{\sqrt{\lambda}} + \sqrt{\lambda}\right) + \eta\frac{|\theta|^{2r+2}}{1 + \sqrt{\lambda}|\theta|^{2r}} \\
&\leq \left(\frac{1}{\sqrt{\lambda}} + \sqrt{\lambda}\right)\sqrt{d}|\theta| + \eta\frac{|\theta|^{2r+2}}{1 + \sqrt{\lambda}|\theta|^{2r}}
\end{aligned} \tag{D.7}$$

and (D.3), it can be shown that

$$\begin{aligned}
&\mathbb{E}\left[|\theta_{n+1}^\lambda|^2\mathbf{1}_{|\theta_n^\lambda|< M_0}\middle|\theta_n^\lambda\right] \\
= \;& \mathbb{E}\left[\left(|\theta_n^\lambda|^2 - 2\lambda\langle\theta_n^\lambda,\, H_{\lambda,c}(\theta_n^\lambda, X_{n+1})\rangle + \lambda^2|H_{\lambda,c}(\theta_n^\lambda, X_{n+1})|^2 + \frac{2\lambda}{\beta}|\xi_{n+1}|^2\right)\mathbf{1}_{|\theta_n^\lambda|< M_0}\middle|\theta_n^\lambda\right] \\
\leq \;& \left(|\theta_n^\lambda|^2 + \frac{2\lambda d}{\beta}\right)\mathbf{1}_{|\theta_n^\lambda|< M_0} + \mathbb{E}\left[\left(2\lambda|\langle\theta_n^\lambda,\, H_{\lambda,c}(\theta_n^\lambda, X_{n+1})\rangle| + \lambda^2|H_{\lambda,c}(\theta_n^\lambda, X_{n+1})|^2\right)\mathbf{1}_{|\theta_n^\lambda|< M_0}\middle|\theta_n^\lambda\right] \\
\leq \;& |\theta_n^\lambda|^2 + \frac{2\lambda d}{\beta} + 2\left(\sqrt{\lambda} + \lambda^{\frac{3}{2}}\right)\sqrt{d}M_0 + 2\eta\sqrt{\lambda}M_0^2 + 4d(\lambda + \lambda^3) + 2\eta^2M_0^2\lambda \\
\leq \;& \left(1 - \frac{\eta\sqrt{\lambda}}{2}\right)|\theta_n^\lambda|^2 + \sqrt{\lambda}\left(\frac{5\eta}{2}M_0^2 + \frac{2\sqrt{\lambda}d}{\beta} + 2(1+\lambda)\sqrt{d}M_0 + 4d(\sqrt{\lambda} + \lambda^{\frac{5}{2}}) + 2\eta^2M_0^2\sqrt{\lambda}\right)
\end{aligned} \tag{D.8}$$

Consequently, (D.6) and (D.8) yield that

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

As a result,

$$\begin{aligned}
\mathbb{E}\left[|\theta_{n+1}^\lambda|^2\right] &\leq \left(1 - \frac{\eta\sqrt{\lambda}}{2}\right)^n\mathbb{E}|\theta_0^\lambda|^2 \\
&+ \sqrt{\lambda}\left(\frac{5\eta}{2}M_0^2 + \frac{2\sqrt{\lambda}d}{\beta} + 2(1+\lambda)\sqrt{d}M_0 + 4d(\sqrt{\lambda} + \lambda^{\frac{5}{2}}) + 2\eta^2M_0^2\sqrt{\lambda}\right)\sum_{j=1}^\infty\left(1 - \frac{\eta\sqrt{\lambda}}{2}\right)^j \\
&\leq \left(1 - \frac{\eta\sqrt{\lambda}}{2}\right)^n\mathbb{E}|\theta_0^\lambda|^2 + \left(5M_0^2 + \frac{4\sqrt{\lambda}d}{\beta\eta} + 4(1+\lambda)\sqrt{d}M_0\eta^{-1} + 8d(\sqrt{\lambda} + \lambda^{\frac{5}{2}})\eta^{-1} + 4\eta M_0^2\sqrt{\lambda}\right).
\end{aligned}$$

□

Proof of Lemma C.2. For any integer $p > 1$, $|\theta_{n+1}^\lambda|^{2p}$ is written as

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

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

$$\begin{aligned}
\mathbb{E}[|\theta_{n+1}^\lambda|^{2p}|\theta_n^\lambda] &= \mathbb{E}\left[\left(|\Delta_n|^2 + \left|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\right|^2 + 2\langle \Delta_n,\, \sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\rangle\right)^p \middle|\theta_n^\lambda\right] \\
&= \sum_{k_1+k_2+k_3=p} \frac{p!}{k_1!k_2!k_3!}\mathbb{E}\left[|\Delta_n|^{2k_1}\left|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\right|^{2k_2}\left(2\langle \Delta_n,\, \sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\rangle\right)^{k_3}\middle|\theta_n^\lambda\right] \\
&\le \mathbb{E}[|\Delta_n|^{2p}|\theta_n^\lambda] + 2p\mathbb{E}\left[|\Delta_n|^{2p-2}\langle \Delta_n,\, \sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\rangle\middle|\theta_n^\lambda\right] \\
&+ \sum_{k=2}^{2p}\binom{2p}{k}\mathbb{E}\left[|\Delta_n|^{2p-k}\left|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\right|^k\middle|\theta_n^\lambda\right] \\
&\le \mathbb{E}[|\Delta_n|^{2p}|\theta_n^\lambda] + \sum_{l=0}^{2(p-1)}\binom{2p}{l+2}\mathbb{E}\left[\left(|\Delta_n|^{2(p-1)-l}\left|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\right|^{q-1}\right)\frac{2\lambda}{\beta}|\xi_{n+1}|^2\middle|\theta_n^\lambda\right] \\
&= \mathbb{E}[|\Delta_n|^{2p}|\theta_n^\lambda] + \binom{2p}{2}\sum_{l=0}^{2(p-1)}\binom{2(p-1)}{l}\mathbb{E}\left[\left(|\Delta_n|^{2(p-1)-l}\left|\sqrt{\frac{2\lambda}{\beta}}\xi_{n+1}\right|^{l}\right)\frac{2\lambda}{\beta}|\xi_{n+1}|^2\middle|\theta_n^\lambda\right] \\
&\le \mathbb{E}[|\Delta_n|^{2p}|\theta_n^\lambda] + 2^{2p-3}p(2p-1)\left(\mathbb{E}[|\Delta_n|^{2p-2}|\theta_n^\lambda]\frac{2\lambda d}{\beta} + \left(\frac{2\lambda}{\beta}\right)^p\mathbb{E}|\xi_{n+1}|^{2p}\right). \qquad \text{(D.9)}
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}\left[|\Delta_n|^{2p}\middle|\theta_n^\lambda\right] &= \sum_{k=0}^{p}\binom{p}{k}|\theta_n^\lambda|^{2(p-k)}\mathbb{E}[r_n^k|\theta_n^\lambda] \\
&= |\theta_n^\lambda|^{2p} + p|\theta_n^\lambda|^{2p-2}\mathbb{E}[r_n|\theta_n^\lambda] + \sum_{k=2}^{p}\binom{p}{k}|\theta_n^\lambda|^{2(p-k)}\mathbb{E}[r_n^k|\theta_n^\lambda]. \qquad \text{(D.10)}
\end{aligned}$$

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

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

and M_0 is defined in the proof of Lemma C.1. We need the following relations to estimate the moments of r_n: for all $x \in \mathbb{R}^d$ and $|\theta| \ge M$,

$$\begin{aligned}
\lambda^2|H_{\lambda,c}(\theta,x)|^2 &\le 4d(\lambda+\lambda^3) + 2\eta^2\lambda|\theta|^2\frac{\lambda|\theta|^{4r}}{(1+\sqrt{\lambda}|\theta|^{2r})^2} \\
&\le 4d\lambda(1+\lambda^2) + 2\eta^2\lambda|\theta|^2 \\
&\le 4d\lambda(1+\lambda^2)|\theta| + 2\eta^2\lambda|\theta|^2 \\
&\le 2\sqrt{\lambda}\eta\left(\frac{2\sqrt{\lambda_{\max}}d(1+\lambda_{\max}^2)}{|\theta|\eta} + \sqrt{\lambda_{\max}}\eta\right)|\theta|^2 \\
&\le 2\sqrt{\lambda}\eta\left(\frac{2d\sqrt{\lambda_{\max}}(1+\lambda_{\max}^2)}{M\eta} + \sqrt{\lambda_{\max}}\eta\right)|\theta|^2 \\
&\le 4\sqrt{\lambda}\eta|\theta|^2 \qquad \text{(D.11)}
\end{aligned}$$

where we have used the inequality (D.3), $0 \le \eta < 1$ and

$$M > \frac{2\sqrt{\lambda_{\max}}d(1+\lambda_{\max}^2)}{(2-\sqrt{\lambda_{\max}}\eta)\eta} \Leftrightarrow \left(\frac{2d\sqrt{\lambda_{\max}}(1+\lambda_{\max}^2)}{M\eta} + \sqrt{\lambda_{\max}}\eta\right) < 2$$

and note that $\frac{2\sqrt{\lambda_{\max}}d(1+\lambda_{\max}^2)}{(2-\sqrt{\lambda_{\max}}\eta)\eta}$ is finite due to $\lambda_{\max}$ being less than $\frac{1}{4\eta^2}$. Moreover, from (D.7), we have the following inequality

$$\begin{aligned}
|2\lambda\langle\theta,\, H_{\lambda,c}(\theta,x)\rangle| &\le 2(\sqrt{\lambda}+\lambda^{1.5})\sqrt{d}|\theta| + 2\eta\sqrt{\lambda}|\theta|^2\frac{\sqrt{\lambda}|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}} \\
&\le 2\sqrt{\lambda}(1+\lambda)\sqrt{d}|\theta| + 2\eta\sqrt{\lambda}|\theta|^2 \\
&\le 2\sqrt{\lambda}\eta\left(\frac{(1+\lambda_{\max})\sqrt{d}}{|\theta|\eta} + \eta\right)|\theta|^2 \\
&\le 2\sqrt{\lambda}\eta\left(\frac{(1+\lambda_{\max})\sqrt{d}}{M\eta} + \eta\right)|\theta|^2 \\
&\le 4\sqrt{\lambda}\eta|\theta|^2
\end{aligned} \tag{D.12}$$

where the last inequality holds since

$$M > \frac{(1+\lambda_{\max})\sqrt{d}}{\eta(2-\eta)} \Leftrightarrow \left(\frac{(1+\lambda_{\max})\sqrt{d}}{M\eta} + \eta\right) \le 2.$$

Thus, r_n^k can be written as

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

Moreover, (D.5) implies that

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

equivalently,

$$p|\theta_n^\lambda|^{2p-2}\mathbb{E}[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}r_n|\theta_n^\lambda] \le -p\frac{\eta\sqrt{\lambda}}{2}|\theta_n^\lambda|^{2p}. \tag{D.13}$$

Using (D.13), the L_{2p}-norm of Δ_n conditional on $\theta_n^\lambda > M$ is given by

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}|\Delta_n|^{2p}\middle|\theta_n^\lambda\right] &\le |\theta_n^\lambda|^{2p} + p|\theta_n^\lambda|^{2p-2}\mathbb{E}[\mathbf{1}_{\{\theta_n^\lambda>M\}}r_n|\theta_n^\lambda] + \sum_{k=2}^{p}\binom{p}{k}|\theta_n^\lambda|^{2(p-k)}\mathbb{E}[\mathbf{1}_{\{\theta_n^\lambda>M\}}|r_n|^k|\theta_n^\lambda] \\
&\le |\theta_n^\lambda|^{2p} - p\frac{\eta\sqrt{\lambda}}{2}|\theta_n^\lambda|^{2p} + \sum_{k=2}^{p}\binom{p}{k}|\theta_n^\lambda|^{2(p-k)}\lambda^{\frac{k}{2}}(8\eta)^k|\theta_n^\lambda|^{2k} \\
&\le |\theta_n^\lambda|^{2p} - p\frac{\eta\sqrt{\lambda}}{2}|\theta_n^\lambda|^{2p} + |\theta_n^\lambda|^{2p}\sum_{k=2}^{p}\binom{p}{k}\lambda^{\frac{k}{2}}(8\eta)^k.
\end{aligned} \tag{D.14}$$

Choose λ such that

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

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 integer $2 \leq k \leq p$. Then, since the following inequality can be obtained

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

we have

$$\mathbb{E}\left[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}|\Delta_n|^{2p}\middle|\theta_n^\lambda\right] \leq (1-\eta\sqrt{\lambda})|\theta_n^\lambda|^{2p} \tag{D.15}$$

and

$$\mathbb{E}\left[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}|\Delta_n|^{2p-2}\middle|\theta_n^\lambda\right] \leq (1-\eta\sqrt{\lambda})|\theta_n^\lambda|^{2(p-2)} \leq \frac{1}{M^2}(1-\eta\sqrt{\lambda})|\theta_n^\lambda|^{2p}. \tag{D.16}$$

By combining (D.9), (D.16) and (D.15), we derive

$$\begin{aligned}\mathbb{E}[\mathbf{1}_{\{|\theta_n^\lambda|>M\}}|\theta_{n+1}^\lambda|^{2p}|\theta_n^\lambda] &\leq (1-\eta\sqrt{\lambda})|\theta_n^\lambda|^{2p}\\ &+ \frac{2^{2p-2}p(2p-1)\lambda d}{M^2\beta}(1-\eta\sqrt{\lambda})|\theta_n^\lambda|^{2p} + 2^{2p-3}p(2p-1)\left(\frac{2\lambda}{\beta}\right)^p\mathbb{E}|\xi_{n+1}|^{2p}\\ &\leq (1-\eta\sqrt{\lambda})\left(1+\frac{2^{2p-2}p(2p-1)\lambda d}{M^2\beta}\right)|\theta_n^\lambda|^{2p}\\ &+ 2^{2p-3}p(2p-1)\left(\frac{2\lambda}{\beta}\right)^p\mathbb{E}|\xi_{n+1}|^{2p}\\ &\leq (1-\eta^2\lambda)|\theta_n^\lambda|^{2p} + 2^{2p-3}p(2p-1)\left(\frac{2\lambda}{\beta}\right)^p\mathbb{E}|\xi_{n+1}|^{2p}\end{aligned} \tag{D.17}$$

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 (D.3)

$$\mathbf{1}_{\{|\theta|\leq M\}}\lambda^2|H_{\lambda,c}(\theta,x)|^2 \leq \lambda\left(4d(1+\lambda_{\max}^2)+2\eta^2M^2\right)$$

and

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

it can be shown that

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}_{\{|\theta_n^\lambda|\leq M\}}|r_n|^k\middle|\theta_n^\lambda\right] &= \mathbb{E}\left[\mathbf{1}_{\{|\theta_n^\lambda|\leq M\}}\left(|2\lambda\langle\theta_n^\lambda,\, H_{\lambda,c}(\theta_n^\lambda, X_{n+1})\rangle| + \lambda^2|H_{\lambda,c}(\theta_n^\lambda, X_{n+1})|^2\right)^k\middle|\theta_n^\lambda\right] \\
&\leq \mathbb{E}\Bigg[\mathbf{1}_{\{|\theta_n^\lambda|\leq M\}}\Bigg(2\lambda\sqrt{M}\sqrt{K(X_{n+1})(1+M^q)+d\sqrt{\lambda}_{\max}+2\eta M^{2r+1}} \\
&+ \lambda\left(4d(1+\lambda_{\max}^2)+2\eta^2M^2\right)\Bigg)^k\Bigg|\theta_n^\lambda\Bigg] \\
&\leq \widetilde{D}_k\lambda^k
\end{aligned}$$

where $\widetilde{D}_k = 2^{k-1}\Big((2\sqrt{M})^k(\mathbb{E}[K(X_0)](1+M^q)+d\sqrt{\lambda}_{\max}+2\eta M^{2r+1})^{k/2}+(4d(1+\lambda_{\max}^2)+$ $2\eta^2M^2)^k\Big)$. 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

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

Consequently, we obtain

$$\begin{aligned}
\mathbb{E}[\mathbf{1}_{\{|\theta_n^\lambda|\leq M\}}|\theta_{n+1}^\lambda|^{2p}|\theta_n^\lambda] &\leq (1-\eta^2\lambda)|\theta_n^\lambda|^{2p}+\eta^2\lambda M^{2p}+\lambda M^{2p}\sum_{k=1}^{p}\binom{p}{k}\lambda^{k-1}\widetilde{D}_k \\
&+ \frac{\lambda d}{\beta}2^{2p-2}p(2p-1)M^{2p-2}\sum_{k=0}^{p-1}\binom{p}{k}\lambda^k\widetilde{D}_k \qquad \text{(D.18)} \\
&+ 2^{2p-3}p(2p-1)\left(\frac{2\lambda}{\beta}\right)^p\mathbb{E}|\xi_{n+1}|^{2p}.
\end{aligned}$$

By defining

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

we conclude that

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

□

Proof of Lemma C.6. We begin by observing that

$$\begin{aligned}
\mathbb{E}|\bar{\theta}^{\lambda}_t - \bar{\zeta}^{\lambda,m}_t|^2 &= -2\lambda \int_{mT}^{t} \mathbb{E}\langle \bar{\zeta}^{\lambda,m}_s - \bar{\theta}^{\lambda}_s, h(\bar{\zeta}^{\lambda,m}_s) - H_{\lambda}(\bar{\theta}^{\lambda}_{\lfloor s \rfloor}, X_{\lceil s \rceil})\rangle ds \\
&= -2\lambda \int_{mT}^{t} \mathbb{E}\langle \bar{\zeta}^{\lambda,m}_s - \bar{\theta}^{\lambda}_s,\, h(\bar{\zeta}^{\lambda,m}_s) - h(\bar{\theta}^{\lambda}_s)\rangle ds \\
&- 2\lambda \int_{mT}^{t} \mathbb{E}\langle \bar{\zeta}^{\lambda,m}_s - \bar{\theta}^{\lambda}_s,\, h(\bar{\theta}^{\lambda}_s) - h(\bar{\theta}^{\lambda}_{\lfloor s \rfloor})\rangle ds \\
&- 2\lambda \int_{mT}^{t} \mathbb{E}\langle \bar{\zeta}^{\lambda,m}_s - \bar{\theta}^{\lambda}_s,\, h(\bar{\theta}^{\lambda}_{\lfloor s \rfloor}) - H(\bar{\theta}^{\lambda}_{\lfloor s \rfloor}, X_{\lceil s \rceil})\rangle ds \\
&- 2\lambda \int_{mT}^{t} \mathbb{E}\langle \bar{\zeta}^{\lambda,m}_s - \bar{\theta}^{\lambda}_s, H(\bar{\theta}^{\lambda}_{\lfloor s \rfloor}, X_{\lceil s \rceil}) - H_{\lambda}(\bar{\theta}^{\lambda}_{\lfloor s \rfloor}, X_{\lceil s \rceil})\rangle ds \\
&\leq 2\lambda a \int_{mT}^{t} \mathbb{E}|\bar{\zeta}^{\lambda,m}_s - \bar{\theta}^{\lambda}_s|^2 ds \\
&+ \frac{\lambda a}{2} \int_{mT}^{t} \mathbb{E}|\bar{\zeta}^{\lambda,m}_s - \bar{\theta}^{\lambda}_s|^2 ds + \int_{mT}^{t} \frac{2\lambda}{a} \mathbb{E}|h(\bar{\theta}^{\lambda}_s) - h(\bar{\theta}^{\lambda}_{\lfloor s \rfloor})|^2 ds \\
&+ \int_{mT}^{t} \Big(- 2\lambda \mathbb{E}\langle \bar{\zeta}^{\lambda,m}_s - \bar{\theta}^{\lambda}_s,\, h(\bar{\theta}^{\lambda}_{\lfloor s \rfloor}) - H(\bar{\theta}^{\lambda}_{\lfloor s \rfloor}, X_{\lceil s \rceil})\rangle \Big) ds \\
&+ \frac{\lambda a}{2} \int_{mT}^{t} \mathbb{E}|\bar{\zeta}^{\lambda,m}_s - \bar{\theta}^{\lambda}_s|^2 ds + \int_{mT}^{t} \frac{2\lambda}{a} \mathbb{E}|H(\bar{\theta}^{\lambda}_{\lfloor s \rfloor}, X_{\lceil s \rceil}) - H_{\lambda,c}(\bar{\theta}^{\lambda}_{\lfloor s \rfloor}, X_{\lceil s \rceil})|^2 ds \\
&= 3\lambda a \int_{mT}^{t} \mathbb{E}|\bar{\zeta}^{\lambda,m}_s - \bar{\theta}^{\lambda}_s|^2 ds + \int_{mT}^{t} A^{\lambda,m}_s + B^{\lambda,m}_s + D^{\lambda,m}_s ds \qquad \text{(D.19)}
\end{aligned}$$

where we have used Proposition B.1 and the Young's inequality in the first inequality and

$$\begin{aligned}
A^{\lambda,m}_t &:= \frac{2\lambda}{a} \mathbb{E}|h(\bar{\theta}^{\lambda}_t) - h(\bar{\theta}^{\lambda}_{\lfloor t \rfloor})|^2 \\
B^{\lambda,m}_t &:= -2\lambda \mathbb{E}\langle \bar{\zeta}^{\lambda,m}_t - \bar{\theta}^{\lambda}_t,\, h(\bar{\theta}^{\lambda}_{\lfloor t \rfloor}) - H(\bar{\theta}^{\lambda}_{\lfloor t \rfloor}, X_{\lceil t \rceil})\rangle \\
D^{\lambda,m}_t &:= \frac{2\lambda}{a} \mathbb{E}|H(\bar{\theta}^{\lambda}_{\lfloor t \rfloor}, X_{\lceil t \rceil}) - H_{\lambda,c}(\bar{\theta}^{\lambda}_{\lfloor t \rfloor}, X_{\lceil t \rceil})|^2.
\end{aligned}$$

In addition, from the definition of $\bar{\theta}^{\lambda}_t$ and (D.3), we have

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

which yields

$$\sqrt{\mathbb{E}|\bar{\theta}^{\lambda}_t - \bar{\theta}^{\lambda}_{\lfloor t \rfloor}|^4} \leq \widetilde{C}_1 \lambda \qquad \text{(D.20)}$$

where $\widetilde{C}_1 = 2^{5/2} \sqrt{8d^2(1+\lambda^4_{\max}) + \eta^4(\mathbb{E}|\theta_0|^4 + A_2/\eta^2) + \frac{3}{\beta^2} d^2}$.

Using Proposition B.2, $A_t^{\lambda,m}$ can be bounded as follows:

$$\begin{aligned}
A_t^{\lambda,m} &\leq \frac{2\lambda}{a} L_X \mathbb{E}[(1+|\bar{\theta}_t^\lambda|+|\bar{\theta}_{\lfloor t \rfloor}^\lambda|)^{2l}|\bar{\theta}_t^\lambda-\bar{\theta}_{\lfloor t \rfloor}^\lambda|^2] \\
&\leq \frac{2\lambda}{a} L_X \sqrt{\mathbb{E}(1+|\bar{\theta}_t^\lambda|+|\bar{\theta}_{\lfloor t \rfloor}^\lambda|)^{4l}}\sqrt{\mathbb{E}|\bar{\theta}_t^\lambda-\bar{\theta}_{\lfloor t \rfloor}^\lambda|^4} \\
&\leq \frac{2\lambda}{a} L_X 3^{2l}\sqrt{(1+\mathbb{E}|\bar{\theta}_t^\lambda|^{4l}+\mathbb{E}|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^{4l})}\sqrt{\mathbb{E}|\bar{\theta}_t^\lambda-\bar{\theta}_{\lfloor t \rfloor}^\lambda|^4} \\
&\leq C_1\lambda^2 \qquad\qquad \text{(D.21)}
\end{aligned}$$

where $L_X = L^2 2^{2\rho-1}(1+\mathbb{E}|X_0|^{2\rho})$ and $C_1 = \frac{2}{a}L_X 9^l\sqrt{(1+2\mathbb{E}|\bar{\theta}_0^\lambda|^{4l}+2\frac{A_{2l}}{\eta^2})}\widetilde{C}_1$ and (D.20) is used for the last inequality.

To estimate $B_t^{\lambda,m}$, we observe that

$$\begin{aligned}
B_t^{\lambda,m} &= -2\lambda\mathbb{E}\langle\bar{\zeta}_t^{\lambda,m}-\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 \\
&\quad -2\lambda\mathbb{E}\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 \\
&\leq -2\lambda\mathbb{E}\left[\mathbb{E}\left[\langle\bar{\zeta}_t^{\lambda,m}-\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,m},\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})ds-\sqrt{\frac{2\lambda}{\beta}}\tilde{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] \\
&\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\sqrt{\mathbb{E}|H_\lambda(\bar{\theta}_{\lfloor t \rfloor}^\lambda, X_{\lceil t \rceil})|^2}\sqrt{\mathbb{E}|h(\bar{\theta}_{\lfloor t \rfloor}^\lambda)-H(\bar{\theta}_{\lfloor t \rfloor}^\lambda, X_{\lceil t \rceil})|^2} \\
&\leq 2\lambda^2\sqrt{6\mathbb{E}|K(X_0)|^2(1+\mathbb{E}|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^{2q})+3\lambda d+3\eta^2\mathbb{E}|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^{4r+2}}\sqrt{4\mathbb{E}|H(\bar{\theta}_{\lfloor t \rfloor}^\lambda, X_{\lceil t \rceil})|^2} \\
&\leq 4\lambda^2\sqrt{6\mathbb{E}|K(X_0)|^2(1+\mathbb{E}|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^{2q})+3\lambda d+3\eta^2\mathbb{E}|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^{4r+2}} \\
&\quad \times\sqrt{4\mathbb{E}|K(X_0)|^2(1+\mathbb{E}|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^{2q})+2\eta^2\mathbb{E}|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^{4r+2}} \\
&\leq C_2\lambda^2 \qquad\qquad \text{(D.22)}
\end{aligned}$$

where

$$\begin{aligned}
C_2 &= 4\sqrt{6\mathbb{E}|K(X_0)|^2(1+\mathbb{E}|\theta_0|^{2q}+\frac{A_q}{\eta^2})+3\lambda d+3\eta^2|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^{4r+2}} \\
&\times \sqrt{4\mathbb{E}|K(X_0)|^2(1+\mathbb{E}|\theta_0|^{2q}+\frac{A_q}{\eta^2})+2\eta^2\left(\mathbb{E}|\bar{\theta}_0^\lambda|^{4r+2}+\frac{A_{2r+1}}{\eta^2}\right)}.
\end{aligned}$$

Note that we have used the independence of $\bar{\theta}_{\lfloor s \rfloor}^\lambda$ and $X_{\lceil s \rceil}$ to obtain the second inequality, and used Remark B.4 and Lemma C.2 to calculate the bound of $\mathbb{E}|H_\lambda(\bar{\theta}_{\lfloor t \rfloor}^\lambda, X_{\lceil t \rceil})|^2$ and $\mathbb{E}|H(\bar{\theta}_{\lfloor t \rfloor}^\lambda, X_{\lceil t \rceil})|^2$.

Moreover, $D_t^{\lambda,m}$ can be estimated as follows, from Remark B.3,

$$\begin{aligned}
D_t^{\lambda,m} &\leq \frac{6\lambda^2}{a}\left[8\mathbb{E}|K(X_0)|^4(1+\mathbb{E}|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^{4q})+d+\eta^2\mathbb{E}|\bar{\theta}_{\lfloor t \rfloor}^\lambda|^{8r+2}\right] \\
&\leq C_3\lambda^2 \qquad\qquad \text{(D.23)}
\end{aligned}$$

where the independence of $\bar{\theta}_{\lfloor s \rfloor}^\lambda$ and $X_{\lceil s \rceil}$ is used and C_3 is given by

$$C_3 = \frac{6}{a}\left[8\mathbb{E}|K(X_0)|^4(1+\mathbb{E}|\bar{\theta}_0^\lambda|^{4q}+A_{2q}/\eta^2)+d+\eta^2(\mathbb{E}|\bar{\theta}_0^\lambda|^{8r+2}+A_{4r+1}/\eta^2)\right].$$

Plugging (D.21), (D.22) and (D.23) into (D.19), one can derive

$$
\begin{aligned}
\mathbb{E}|\bar{\theta}_t^{\lambda} - \bar{\zeta}_t^{\lambda,m}|^2 &\leq 3\lambda a \int_{mT}^{t} \mathbb{E}|\bar{\theta}_s^{\lambda} - \bar{\zeta}_s^{\lambda,m}|^2 ds + \int_{nT}^{t} (C_1 + C_2 + C_3)\lambda^2 ds \\
&\leq 3\lambda a \int_{mT}^{t} \mathbb{E}|\bar{\theta}_s^{\lambda} - \bar{\zeta}_s^{\lambda,m}|^2 ds + (C_1 + C_2 + C_3)\lambda < \infty
\end{aligned}
$$

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

$$
\mathbb{E}|\bar{\theta}_t^{\lambda} - \bar{\zeta}_t^{\lambda,m}|^2 \leq c\lambda
$$

where $c = e^{3a}(C_1 + C_2 + C_3)$. □

Proof of Lemma C.7. Since $Z_t^{\lambda} = \bar{\zeta}_t^{\lambda,0}$ and $t \in [mT, (m+1)T)$, we can write

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

where we have used the fact $W_1(\mu, \nu) \leq w_{1,2}(\mu, \nu)$ for $\mu, \nu \in \mathcal{P}_{V_2}$ for the second inequality. Using Remark C.5 and $\lambda(t - kT) \geq m - k$, we further have

$$
\begin{aligned}
w_{1,2}\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,k}\right), \mathcal{L}\left(\bar{\zeta}_t^{\lambda,k-1}\right)\right) &\leq \hat{c}e^{-\dot{c}\lambda(t-kT)} w_{1,2}\left(\mathcal{L}(\bar{\theta}_{kT}^{\lambda}), \mathcal{L}(\bar{\zeta}_{kT}^{\lambda,k-1})\right) \\
&\leq \hat{c}e^{-\dot{c}(m-k)} w_{1,2}\left(\mathcal{L}(\bar{\theta}_{kT}^{\lambda}), \mathcal{L}(\bar{\zeta}_{kT}^{\lambda,k-1})\right) \\
&\leq \hat{c}e^{-\dot{c}(m-k)} W_2\left(\mathcal{L}(\bar{\theta}_{kT}^{\lambda}), \mathcal{L}(\bar{\zeta}_{kT}^{\lambda,k-1})\right) \sqrt{\mathbb{E}\left|1 + V_2(\bar{\theta}_{kT}^{\lambda}) + V_2(\bar{\zeta}_{kT}^{\lambda,k-1})\right|^2} \\
&\leq \hat{c}e^{-\dot{c}(m-k)} W_2\left(\mathcal{L}(\bar{\theta}_{kT}^{\lambda}), \mathcal{L}(\bar{\zeta}_{kT}^{\lambda,k-1})\right) \\
&\times \left[1 + \sqrt{\mathbb{E}[V_4(\bar{\theta}_{kT}^{\lambda})]} + \sqrt{\mathbb{E}[V_4(\bar{\zeta}_{kT}^{\lambda,k-1})]}\right] \\
&\leq \hat{c}e^{-\dot{c}(m-k)} \sqrt{\lambda} \sqrt{e^{3a}(C_1 + C_2 + C_3)} \left[1 + \sqrt{2\mathbb{E}|\theta_0|^4 + 2 + 2\frac{A_2}{\eta^2}}\right. \\
&+ \left.\sqrt{2\mathbb{E}|\theta_0|^4 + 2 + 2\frac{A_2}{\eta^2} + \frac{\tilde{c}(4)}{\bar{c}(4)}}\right]
\end{aligned} \tag{D.24}
$$

where the Cauchy-Schwarz inequality is applied to the third inequality and Lemma C.3, C.6 and C.4 are used for the last inequality. By combining the two inequalities above, we obtain

$$
\begin{aligned}
W_1\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,m}\right), \mathcal{L}\left(Z_t^{\lambda}\right)\right) &\leq \hat{c}\sqrt{\lambda}\sqrt{e^{3a}(C_1 + C_2 + C_3)} \left[1 + \sqrt{2\mathbb{E}|\theta_0|^4 + 2 + 2\frac{A_2}{\eta^2}}\right. \\
&+ \left.\sqrt{2\mathbb{E}|\theta_0|^4 + 2 + 2\frac{A_2}{\eta^2} + \frac{\tilde{c}(4)}{\bar{c}(4)}}\right] \sum_{k=1}^{m} e^{-\dot{c}(m-k)} \\
&\leq z_1\sqrt{\lambda}
\end{aligned}
$$

where $z_1 = \frac{\hat{c}}{1 - \exp(-\dot{c})} \sqrt{e^{3a}(C_1 + C_2 + C_3)} \left[1 + \sqrt{2\mathbb{E}|\theta_0|^4 + 2 + 2\frac{A_2}{\eta^2}} + \sqrt{2\mathbb{E}|\theta_0|^4 + 2 + 2\frac{A_2}{\eta^2} + \frac{\tilde{c}(4)}{\bar{c}(4)}}\right]$. □

Proof of Lemma C.8. We begin by observing that

$$\begin{aligned}
W_2\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,k}\right),\mathcal{L}\left(\bar{\zeta}_t^{\lambda,k-1}\right)\right) &\leq \sqrt{2w_{1,2}\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,k}\right),\mathcal{L}\left(\bar{\zeta}_t^{\lambda,k-1}\right)\right)} \\
&\leq \lambda^{1/4}e^{-\dot{c}(m-k)/2}\left[\hat{c}\sqrt{e^{3a}(C_1+C_2+C_3)}\left(1+\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}}\right.\right. \\
&+ \left.\left.\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}+\frac{\tilde{c}(4)}{\bar{c}(4)}}\right)\right]^{1/2}
\end{aligned}$$

where we have used the fact $W_2 \leq \sqrt{2w_{1,2}}$ for the first inequality, and the second inequality follows from (D.24). Consequently, we derive

$$\begin{aligned}
W_2\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,m}\right),\mathcal{L}\left(Z_t^{\lambda}\right)\right) &\leq \sum_{k=1}^{m} W_2\left(\mathcal{L}\left(\bar{\zeta}_t^{\lambda,k}\right),\mathcal{L}\left(\bar{\zeta}_t^{\lambda,k-1}\right)\right) \\
&\leq \lambda^{1/4}\left[\hat{c}\sqrt{e^{3a}(C_1+C_2+C_3)}\left(1+\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}}\right.\right. \\
&+ \left.\left.\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}+\frac{\tilde{c}(4)}{\bar{c}(4)}}\right)\right]^{1/2}\sum_{k=1}^{m}e^{-\dot{c}(m-k)/2} \\
&\leq \lambda^{1/4}z_2
\end{aligned}$$

where

$$z_2 = \frac{\sqrt{\hat{c}}e^{3a/4}(C_1+C_2+C_3)^{1/4}}{1-\exp(-\dot{c}/2)}\left[\left(1+\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}}+\sqrt{2\mathbb{E}|\theta_0|^4+2+2\frac{A_2}{\eta^2}+\frac{\tilde{c}(4)}{\bar{c}(4)}}\right)\right]^{1/2}.$$

□

E Details of Experiments

E.1 Image Classification

The experiments are exactly replicated in the official implementation of Zhuang et al. (2020). More specifically, VGG11, ResNet34 and DenseNet121 are trained for 500 epochs. We apply a weight deacy of 0.0005 and decay the initial learning rate by 10 after 150 epochs to all optimizers. Batch normalization proposed in Ioffe & Szegedy (2015) is employed to prevent the models from overfitting and boost the training speed for all three models. The batch size is 128.

Regarding hyperparameter values of Adam, AdaBelief, AdamP, AdaBound, AMSGrad and RMSProp, the best hyperparameters are used across all the experiments in Luo et al. (2019) and Zhuang et al. (2020), which are $\lambda = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-8}$. For SGD, we set the momentum to 0.9 for SGD. For THεO POULA, the best hyperparameters are $\lambda = 0.1$, $\varepsilon = 0.1$ and $\beta = 10^{12}$.

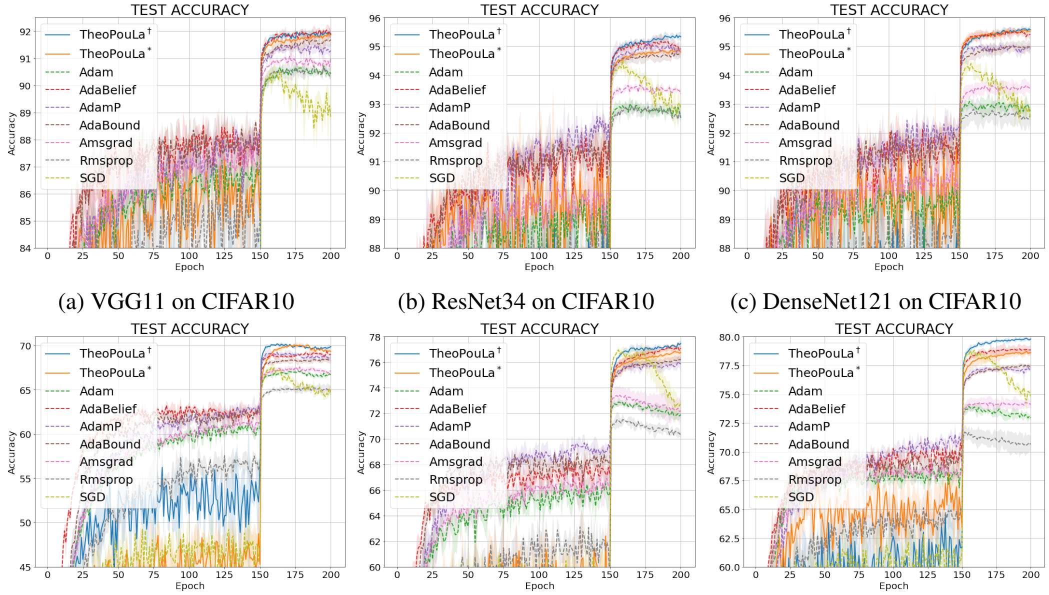
Figure 2 shows test accuracy for VGG11, ResNet34 and DenseNet121 on CIFAR-10 and CIFAR-100.

E.2 Language Modeling

We conduct language modeling for Penn Treebank (PTB) data set. For this task, we train the AWD LSTM of Merity et al. (2018) for 750 epochs. The details of models can be found in the official implementation of AWD-LSTM [1].

For NT-ASGD and averaged THεO POULA, the constant learning rate of 30 is used for 2 and 3-layer LSTMs. For 1-layer LSTMs, we set to 10. $\varepsilon = 100$ and $\beta = 10^{10}$ are set across all the experiments.

[1] https://github.com/salesforce/awd-lstm-lm



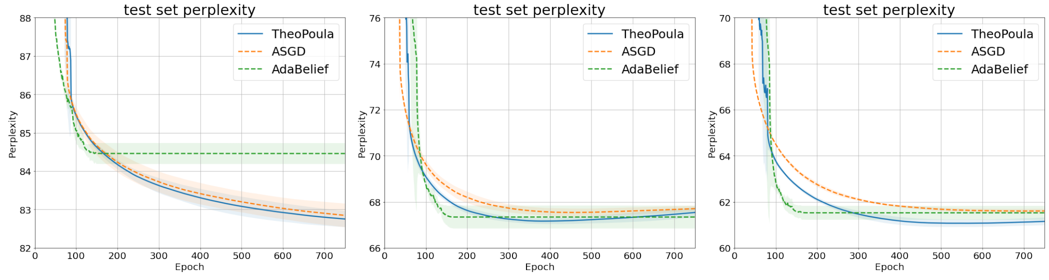
(a) VGG11 on CIFAR10 (b) ResNet34 on CIFAR10 (c) DenseNet121 on CIFAR10

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

Figure 2: Test accuracy for VGG11, ResNet34 and DenseNet121 on CIFAR-10 and CIFAR-100. THεOPOULA† and THεOPOULA* represent the performances of THεO POULA under the best and second best hyperparameters, respectively.

For AdaBelief, we used the best hyperparameters reported in Zhuang et al. (2020). Thus, we obtain the same results of AdaBelief in Zhuang et al. (2020). Specifically, we set $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-12}$ and an initial learning rate of 0.01 for 2 and 3-layer LSTMs. $\lambda = 0.001$ and $\varepsilon = 10^{-16}$ are used for 1-layer LSTMs.

The averaging is triggered when no improvement has been made for 5 consecutive epochs for THεO POULA. Also, 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 have searched the optimal hyperparameters for the development-based learning rate decay among $\delta = \{0.1, 0.5\}$ and $k = \{5, 10, 20\}$. We have found $\delta = 0.1$ and $k = 5$ yield the best performance. Figure 3 displays test perplexity for different AWD-LSTM models on PTB.



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

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

E.3 Effectiveness of the boosting function

This subsection empirically tests the effectiveness of the boosting function in our algorithm. THεO POULA without the boosting function updates the parameter as follows:

$$\theta_0^\lambda := \theta_0, \qquad \theta_{n+1}^\lambda := \theta_n^\lambda - \lambda H_{\lambda,c}(\theta_n^\lambda, X_{n+1}) + \sqrt{2\lambda\beta^{-1}}\xi_{n+1}, \quad n \in \mathbb{N},$$

where $H_{\lambda,c} := (H^{(1)}_{\lambda,c}(\theta,x),\cdots,H^{(d)}_{\lambda,c}(\theta,x))^T$ is given by

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

and $\{\xi_n\}_{n\geq 1}$ is a sequence of independent standard d-dimensional Gaussian random variables.

Indeed, this is a special case of THεO POULA with $\varepsilon = \infty$. We train VGG11, ResNet34 and DenseNet121 on CIFAR-10 and CIFAR-100 using the iterating rule of (E.1). The hyperparameters are the same with the best hyperparameters of THεO POULA. As Table 4 shows, THεO POULA without the boosting function is worsen than THεO POULA. The result indicates that the addition of the boosting function leads to a meaningful increase of test accuracy, validating that the boosting function works well for the sparsity of neural networks as expected.

Table 4: The best accuracy for VGG11, ResNet34 and DenseNet121 on CIFAR-10 and CIFAR-100. THεOPOULA† represents the performances of THεO POULA with the best hyperparameters. THεOPOULA$(\varepsilon = \infty)$ means the performance without the boosting function with the same hyperparameters.

dataset	CIFAR-10			CIFAR-100		
model	VGG	ResNet	DenseNet	VGG	ResNet	DenseNet
THεOPOULA†	**92.10**	**95.43**	**95.66**	**70.31**	**77.53**	**79.90**
THεOPOULA$(\varepsilon = \infty)$	91.48	94.93	95.26	68.11	75.91	77.99

F Additional experiments

F.1 Effect of β on the performance of THεO POULA

This subsection examines the effect of β on the performance of THεO POULA. We conduct experiments for VGG11 and ResNet34 on CIFAR10 and CIFAR100 with different values of β ranging from 10^4 to 10^{12}. As shown in Table 5, THεO POULA achieves the highest accuracy when β is $10^8 \sim 10^{12}$, which is consistent with the phenomenon, so called the cold posterior effect, see Wenzel et al. (2020) and Aitchison (2021).

Table 5: The accuracy for VGG11 and ResNet34 on CIFAR-10 and CIFAR-100. We use the best hyperparameters for λ and ε in Appendix E.1.

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

F.2 Experiments with $\eta \neq 0$

We conduct additional experiments with $\eta \neq 0$ to demonstrate that there is no gap between theory and practice of our work. When the regularization parameter r is large (possibly, overestimated) and the dimension of θ is big, $|\theta|^{2r}$ becomes substantially huge. As a result, the stochastic gradient of the regularization term, $\eta\frac{\theta^{(i)}|\theta|^{2r}}{1+\sqrt{\lambda}|\theta|^{2r}}$, in (8) will approximately behave like $\frac{\eta}{\sqrt{\lambda}}\theta^{(i)}$, which is equivalent to ℓ_2-regularization. In all the numerical experiments in Table 2, we applied a weight decay with 5×10^{-4} for image classification and with 1.2×10^{-6} for language modeling. That is, by choosing $\eta = 5 \times 10^{-4}\sqrt{\lambda}$ and large r, one can obtain accuracy of models with ℓ_2-regularization.

Table 6 shows that the accuracy for VGG, ResNet and DenseNet on CIFAR-10 and CIFAR-100 with $r = 10$ and $\eta = 5 \times 10^{-4}\sqrt{\lambda}$. One observes a very similar performance by THεO POULA as in Table 2 without any noticeable loss of accuracy.

Table 6: The accuracy for VGG11, ResNet34 and DenseNet121 on CIFAR-10 and CIFAR-100. We use the best hyperparameters reported in Appendix E.1 with $r = 10$ and $\eta = 5 \times 10^{-4}\sqrt{\lambda}$.

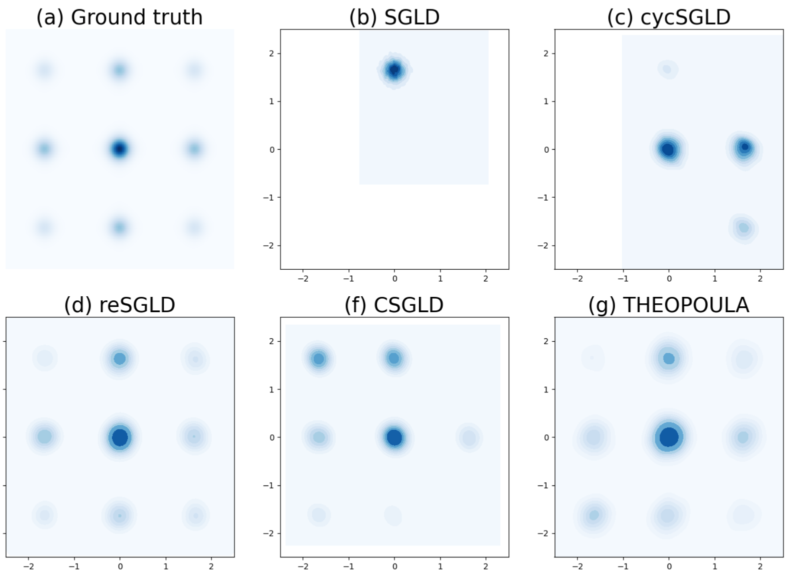
dataset	CIFAR-10			CIFAR-100		
model	VGG	ResNet	DenseNet	VGG	ResNet	DenseNet
THεO POULA ($\eta \neq 0$)	92.2	95.38	95.69	70.07	77.78	80.47

F.3 Sampling from a synthetic multi-modal distribution

We provide a brief overview of recent progress on sampling and Bayesian neural networks to describe the potential of THεO POULA as a sampling algorithm. Deng et al. (2020b) proposed an adaptive MCMC algorithm, called CSGLD, which uses a scalable dynamic importance sampler to flatten the target distribution and reduce the energy barriers to escape local optima. Deng et al. (2020a) devloped reSGMCMC motivated by replica exchange monte carlo algorithm. In particular, reSGMCMC obtained the state-of-the art results on CIFAR-10, CIFAR-100 and SVHN in Bayesian neural networks. Zhang et al. (2020) developed cyclical stochastic gradient MCMC with a cyclical stepsize schedule.

We replicate the simulation of a synthetic multi-modal distribution in Deng et al. (2020b) to evaluate the performance of THεO POULA as a sampling method. A target distribution is $\pi(\mathbf{x}) \propto e^{-U(\mathbf{x})}$ where $U(\mathbf{x}) = \sum_{i=1}^{2} \frac{x(i)^2 - 10\cos(1.2\pi x(i))}{3}$ and $\mathbf{x} = (x(1), x(2))$. Detail of the setting in the experiment such as hyperparameters, regularizer, training epochs can be found in Appendix D.3 of Deng et al. (2020b). THεO POULA is compared with SGLD, CSGLD (Deng et al. (2020b)), reSGLD (Deng et al. (2020a)), cycSGLD (Zhang et al. (2020)). For THεO POULA, we used $\lambda = 0.05$, $\varepsilon = 1$ and $T = 0.3$. A resampling scheme is used for CSGLD. Figure 4 illustrates that THεO POULA recovers the target multi-modal distribution successfully without the local trap issue observed in SGLD and cycSGLD.

Figure 4: simulations of a multi-modal distribution.



G Table of Constants

Table 7 displays full expressions for constants which appear in the main results of this paper. In addition, Table 8 shows all main constants and their dependencies on key parameters such as d, β, the moments of $K(X_0)$ and η.

Table 7: Explicit expression for constants with $\hat{c}$ and $\dot{c}$ from Proposition 3.14 of Chau et al. (2019).

Symbol	Full Expression
M	$\max\left\{M_0, 1, \frac{2\sqrt{\lambda_{\max}}d(1+\lambda_{\max}^2)}{(2-\sqrt{\lambda_{\max}}\eta)\eta}, \frac{(1+\lambda_{\max})\sqrt{d}}{\eta(2-\eta)}, \frac{2^{2p-2}p(2p-1)d}{\eta\beta}\right\}$
$\widetilde{D}_k$	$2^{k-1}\Big[(2\lambda_{\max}\sqrt{M})^k(\mathbb{E}[K(X_0)](1+M^q)+d\sqrt{\lambda}_{\max}+2\eta M^{2r+1})^{k/2}$ $+(4d(1+\lambda_{\max}^2)+2\eta^2M^2)^k\Big], \quad k=1,\cdots,8(2r+1)$
A_p	$\eta^2M^{2p}+M^{2p}\sum_{k=1}^{p}\binom{p}{k}\lambda_{\max}^{k-1}\widetilde{D}_k$ $+2^{2p-3}p(2p-1)\left(\frac{2dM^{2p-2}}{\beta}\sum_{k=0}^{p-1}\binom{p}{k}\lambda_{\max}^{k}\widetilde{D}_k+\frac{2}{\beta}\left(\frac{2\lambda_{\max}}{\beta}\right)^{p-1}d^p(2p-1)!!\right),$ FOR $p=1,\cdots,8(2r+1)$
$\overline{M}_p$	$\sqrt{\frac{1}{3}+4B/(3A)+4d/(3A\beta)+4(p-2)/(3A\beta)}$
$\overline{c}(p)$	$\frac{Ap}{4}, \quad p=1,\cdots,8(2r+1)$
$\tilde{c}(p)$	$\frac{3}{4}Apv_p(\overline{M}_p), \quad p=1,\cdots,8(2r+1)$
C_1	$\frac{L^2 2^{2\rho+5/2}3^{2l}}{a}(1+\mathbb{E}\vert X_0\vert^{2\rho})\sqrt{(1+2\mathbb{E}\vert\bar{\theta}_0^\lambda\vert^{4l}+2\frac{A_{2l}}{\eta^2})}$ $\times\sqrt{8d^2(1+\lambda_{\max}^4)+\eta^4(\mathbb{E}\vert\theta_0\vert^4+A_2/\eta^2)+\frac{3}{\beta^2}d^2}$
C_2	$4\sqrt{6\mathbb{E}\vert K(X_0)\vert^2(1+\mathbb{E}\vert\theta_0\vert^{2q}+\frac{A_q}{\eta^2})+3\lambda_{\max}d+3\eta^2\vert\bar{\theta}_{\lfloor t\rfloor}^\lambda\vert^{4r+2}}$ $\times\sqrt{4\mathbb{E}\vert K(X_0)\vert^2(1+\mathbb{E}\vert\theta_0\vert^{2q}+\frac{A_q}{\eta^2})+2\eta^2\left(\mathbb{E}\vert\bar{\theta}_0^\lambda\vert^{4r+2}+\frac{A_{2r+1}}{\eta^2}\right)}$
C_3	$\frac{6}{a}\left[8\mathbb{E}\vert K(X_0)\vert^4(1+\mathbb{E}\vert\bar{\theta}_0^\lambda\vert^{4q}+A_{2q}/\eta^2)+d+\eta^2(\mathbb{E}\vert\bar{\theta}_0^\lambda\vert^{8r+2}+A_{4r+1}/\eta^2)\right].$
z_1	$\frac{\hat{c}\sqrt{e^{3a}(C_1+C_2+C_3)}}{1-\exp(-\dot{c})}\left[1+\sqrt{2\mathbb{E}\vert\theta_0\vert^4+2+2\frac{A_2}{\eta^2}}+\sqrt{2\mathbb{E}\vert\theta_0\vert^4+2+2\frac{A_2}{\eta^2}+\frac{\tilde{c}(4)}{\overline{c}(4)}}\right]$
z_2	$\frac{\sqrt{\hat{c}}e^{3a/4}(C_1+C_2+C_3)^{1/4}}{1-\exp(-\dot{c}/2)}\left(1+\sqrt{2\mathbb{E}\vert\theta_0\vert^4+2+2\frac{A_2}{\eta^2}}+\sqrt{2\mathbb{E}\vert\theta_0\vert^4+2+2\frac{A_2}{\eta^2}+\frac{\tilde{c}(4)}{\overline{c}(4)}}\right)^{\frac{1}{2}}$
a_1	$2^l(\mathbb{E}[K(X_0)]+\eta)$
R_0	$R_0 := \max\{\sqrt{B/A}, \sqrt{2d/(\beta L\mathbb{E}(1+\vert X_0\vert)^\rho)}\}$
K	$L\mathbb{E}[(1+\vert X_0\vert)^\rho](1+4R_0)^l$
M_1	$(z_1+\sqrt{e^{3a}(C_1+C_2+C_3)})$
M_2	$\hat{c}\left[1+\mathbb{E}\left[V_2\left(\theta_0\right)\right]+\int_{\mathbb{R}^d}V_2(\theta)\pi_\beta(d\theta)\right]$
M_3	$\sqrt{e^{3a}(C_1+C_2+C_3)}$
M_4	$\sqrt{2\hat{c}\left(1+\mathbb{E}\left[V_2\left(\theta_0\right)\right]+\int_{\mathbb{R}^d}V_2(\theta)\pi_\beta(d\theta)\right)}$
M_5	$\left(\frac{a_1}{l+1}\sqrt{\mathbb{E}\vert\theta_0\vert^{2l}+\frac{A_l}{\eta^2}}+\frac{a_1}{l+1}\sqrt{\mathbb{E}\vert\theta_\infty\vert^{2l}}+2\mathbb{E}[K(X_0)]\right)$
M_6	$\left[\frac{d}{2}\log\left(\frac{eK}{A}\left(\frac{B\beta}{d}+1\right)\right)-\log\left(1-e^{-(R_0\sqrt{K\beta}-\sqrt{d})^2}\right)\right]$

Table 8: Main constants and their dependency to key parameters

CONSTANT	KEY PARAMETERS			
	d	β	MOMENTS OF X_0	η
A	-	-	$\mathcal{O}(\mathbb{E}K(X_0))$	-
B	-	-	$\mathcal{O}(\mathbb{E}K(X_0)^{q+2})$	$\mathcal{O}(\frac{1}{\eta^{q+1}})$
R	-	-	$\mathcal{O}(\mathbb{E}\lvert X_0\rvert^{\rho})$	$\mathcal{O}(\frac{1}{\eta^{2r-q}})$
a	-	-	$\mathcal{O}\left(\mathbb{E}\lvert X_0\rvert^{\rho(q-1)}\right)$	$\mathcal{O}(\frac{1}{\eta^{(2r-q)(q-1)}})$
A_p	$poly(d)$	$\mathcal{O}(\frac{d}{\beta})$	$\mathcal{O}(\mathbb{E}K\lvert X_0\rvert^{p/2})$	$\mathcal{O}(\frac{1}{\eta^{p-2}})$
$\dot{c}$	$\mathcal{O}(e^{-d})$	INHERITED FROM CONTRACTION ESTIMATES IN EBERLE ET AL. (2019)		
$\hat{c}$	$\mathcal{O}(e^{d})$	INHERITED FROM CONTRACTION ESTIMATES IN EBERLE ET AL. (2019)		