Taming the Interacting Particle Langevin Algorithm: The Superlinear case[*]

Tim Johnston[3], Nikolaos Makras[1], and Sotirios Sabanis[1,2,3,4]

[1]*School of Mathematics, University of Edinburgh, UK*
[2]*National Technical University of Athens, Greece*
[3]*Université Paris Dauphine-PSL, France*
[4]*Alan Turing Institute, UK*
[5]*Archimedes/Athena Research Centre, Greece*

June 3, 2025

Abstract

Recent advances in stochastic optimization have yielded the interacting particle Langevin algorithm (IPLA), which leverages the notion of interacting particle systems (IPS) to efficiently sample from approximate posterior densities. This becomes particularly crucial in relation to the framework of Expectation-Maximization (EM), where the E-step is computationally challenging or even intractable. Although prior research has focused on scenarios involving convex cases with gradients of log densities that grow at most linearly, our work extends this framework to include polynomial growth. Taming techniques are employed to produce an explicit discretization scheme that yields a new class of stable, under such non-linearities, algorithms which are called tamed interacting particle Langevin algorithms (tIPLA). We obtain non-asymptotic convergence error estimates in Wasserstein-2 distance for the new class under the best known rate.

1. Introduction

The Expectation-Maximization (EM) algorithm is widely used for approximating maximisers of posterior distributions. Applications span, but are not confined to, hyperparameter estimation, mixture models, hidden variable models, missing or incomplete data and variational inference [4]. At its essence, each iteration of the EM algorithm consists of two fundamental steps: the Expectation (E) and the Maximization (M) step. The algorithm is defined by alternating iteratively between these two steps. Given a data specification $p_\theta(x, y)$ parameterized by θ, where x represents the latent variable (interpreted often as incomplete data) and y the observed data, our aim is to find θ^* that maximizes the marginal likelihood $q_\theta(y) = \int_{\mathbb{R}^{d^x}} p_\theta(x, y) dx$.

When the integral in the E-step is computationally challenging or intractable, Markov Chain Monte Carlo (MCMC) methods are often employed, traditionally using Metropolis-Hastings-type algorithms [16, 23]. However, those MCEM (Monte Carlo Expectation-Maximization) methods have two considerable limitations. Firstly, the underlying EM algorithm is not guaranteed to converge towards a global maximum and is susceptible to local mode entrapment, as documented in [4]. Lastly, in the case of the utilization of MALA (Metropolis-Hastings Adjusted Langevin Algorithm), it is well known that scalability issues arise, particularly in high-dimensional settings.

In this landscape, a new method was introduced by [3], such that samples of the latent space variable were generated via an Unadjusted Langevin Algorithm (ULA) chain. An alternative avenue was pioneered in [10], in which the study of the limiting behaviour of various gradient flows associated with appropriate free energy functionals led to an interacting particle system (IPS) that provides efficient estimates for maximum likelihood estimations. Further, [1] expanded on this framework by injecting noise into the dynamics of the parameter θ itself, thereby transitioning from deterministic to stochastic dynamics. This key modification provides a stochastic system with an invariant measure, allowing the establishment of non-asymptotic convergence estimates for the resulting algorithm that targets θ^*, the maximiser of the marginal likelihood.

In our work, while the convexity assumption is maintained, we address the challenge posed by superlinear growth exhibited by gradients of log densities, which makes other known algorithms, such as vanilla Langevin based algorithms, unstable. To counteract this, we implement taming techniques, initially researched for non-globally Lipschitz drifts for SDEs in [8] and subsequently in [18] and [19]. The latter approach has found applications in optimization and machine

---

[*]To appear in *Applied Mathematics & Optimization* (Springer).

learning and led to the design of new MCMC algorithms as one typically deals with high nonlinear objective function, see e.g. [2, 12, 14, 20]. The underlying principle of all these algorithms is the rescaling of ULA's drift coefficient in such a way that maintains stability without significantly increasing computational complexity as in the case of implicit schemes, or by introducing additional constrains via adaptive stepsizes.

In this paper, we study two new algorithms from the tamed interacting particle Langevin algorithms (tIPLA) class, namely the coordinate wise version, known as tIPLAc, and the uniformly tamed version tIPLAu. Those algorithms are tamed versions of IPLA (developed in [1]) as explained in Subsection 2.2. The best known rate of convergence is recovered for both algorithms and our estimates are explicit regarding their dependencies on dimension and number of particles employed, which are denoted by d and N respectively.

1.1. Notation

We conclude this subsection with some basic notation. For $u, v \in \mathbb{R}^d$, define the scalar product $\langle u, v \rangle = \sum_{i=1}^{d} u_i v_i$ and the Euclidian norm $|u| = \langle u, u \rangle^{1/2}$. For all continuously differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$, ∇f denotes the gradient. The integer part of a real number x is denoted by $\lfloor x \rfloor$. We denote by $\mathcal{P}(\mathbb{R}^d)$ the set of probability measures on $\mathcal{B}(\mathbb{R}^d)$ and for any $p \in \mathbb{N}$, $\mathcal{P}_p(\mathbb{R}^d) = \{\pi \in \mathcal{P} : \int_{\mathbb{R}^d} |x|_p^p d\pi(x) < \infty\}$ denotes the set of all probability measures over $\mathcal{B}(\mathbb{R}^d)$ with finite p-th moment. For any two Borel probability measures μ and ν, we define the Wasserstein distance of order $p \geq 1$ as

$$W_p(\mu, \nu) = \left(\inf_{\zeta \in \prod(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^p d\zeta(x, y) \right)^{1/p},$$

where $\prod(\mu, \nu)$ is the set of all couplings of μ and ν. Moreover, for any $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, there exists a coupling $\zeta^* \in \prod(\mu, \nu)$ such that for any coupling (X, Y) distributed according to ζ^*, $W_p(\mu, \nu) = \mathbb{E}^{1/p}[|X - Y|^p]$.

2. Setting and Definitions

2.1. Initial setup

Let $p_\theta(x, y)$ be the aforementioned joint probability density function of the latent variable x and for fixed (observed) data y. The goal of maximum marginal likelihood estimation (MMLE) is to find the parameter θ^* that maximises the marginal likelihood, see [4]. To address the log-density formulation, we define the negative log-likelihood for fixed $y \in \mathbb{R}^{d^y}$ as

$$U_y(\theta, x) := -\log p_\theta(x, y).$$

Thus we arrive at the following notation for the quantity we aim to maximize

$$k_y(\theta) = q_\theta(y) = \int_{\mathbb{R}^{dx}} p_\theta(x, y) dx = \int_{\mathbb{R}^{dx}} \exp(-U_y(\theta, x)) dx.$$

Henceforth, for notational brevity, we denote $h_y(v) = \nabla U_y(v)$ so that $h_y^x(v) = \nabla_x U_y(v)$ and $h_y^\theta(v) = \nabla_\theta U_y(v)$, with $v := (\theta, x)$. Additionally, we drop the reference to the (fixed) data y. To summarize, the optimization problem can be described as

$$\text{maximize } \mathbb{R}^{d^\theta} \ni \theta \to k(\theta) = \int_{\mathbb{R}^{dx}} \exp(-U(\theta, x)) dx.$$

Following the convention of [10], N particles $\mathcal{X}_t^{i,c,N}$ for $i \in \{1, \ldots, N\}$ are used to estimate the gradient of $k(\theta)$, which are governed by the continuous-time dynamics as given below

$$d\vartheta_t^{c,N} = -\frac{1}{N} \sum_{j=1}^{N} \nabla_\theta U(\vartheta_t^{c,N}, \mathcal{X}_t^{j,c,N}) dt + \sqrt{\frac{2}{N}} dB_t^{0,N}, \tag{1}$$

$$d\mathcal{X}_t^{i,c,N} = -\nabla_x U(\vartheta_t^{c,N}, \mathcal{X}_t^{i,c,N}) dt + \sqrt{2} dB_t^{i,N}, \text{ for } i = 1, \ldots, N, \tag{2}$$

where $\{(B_t^{i,N})_{t \geq 0}\}_{0 \leq i \leq N}$ is a family of independent Brownian motions. The discrete time Markov chain associated with the above IPS in equations (1)-(2) is obtained by the corresponding Euler-Maruyama discretization scheme of the Langevin

SDEs (1)-(2)

$$\theta_0^{\lambda_{\text{EM}},c} = \theta_0, \ \theta_{n+1}^{\lambda_{\text{EM}},c} = \theta_n^{\lambda_{\text{EM}},c} - \frac{\lambda}{N}\sum_{i=1}^{N} h^\theta(\theta_n^{\lambda_{\text{EM}},c}, X_n^{i,\lambda_{\text{EM}},c}) + \sqrt{\frac{2\lambda}{N}}\xi_{n+1}^{(0)},$$

$$X_0^{i,\lambda_{\text{EM}},c} = x_0^i, \ X_{n+1}^{i,\lambda_{\text{EM}},c} = X_n^{i,\lambda_{\text{EM}},c} - \lambda h^x(\theta_n^{\lambda_{\text{EM}},c}, X_n^{i,\lambda_{\text{EM}},c}) + \sqrt{2\lambda}\xi_{n+1}^{(i)},$$

for $i = 1, \ldots, N$, where $\theta_0 \in \mathbb{R}^{d^\theta}$, $x_0^i \in \mathbb{R}^{d^x}$, $\lambda > 0$ is the step-size parameter. Moreover, $(\xi_n^{(0)})_{n \in \mathbb{N}}$ and $(\xi_n^{(i)})_{n \in \mathbb{N}}$, $\forall i \in \{1, \ldots, N\}$, are sequences of i.i.d. standard $d^\theta-$ and d^x-dimensional Gaussian variables respectively. A time-scaled version of the original equations (1)-(2) is also introduced, playing a central role in the proof strategy as explained in Subsection 3.5.

$$d\vartheta_t^{u,N} = -\frac{1}{N^{p+1}}\sum_{j=1}^{N}\nabla_\theta U(\vartheta_t^{u,N}, \mathcal{X}_t^{j,u,N})dt + \sqrt{\frac{2}{N^{p+1}}}dB_t^{0,N}, \tag{3}$$

$$d\mathcal{X}_t^{i,u,N} = -\frac{1}{N^p}\nabla_x U(\vartheta_t^{u,N}, \mathcal{X}_t^{i,u,N})dt + \sqrt{\frac{2}{N^p}}dB_t^{i,N}, \ \text{ for } i = 1, \ldots, N, \tag{4}$$

where $p = 2\ell + 1$ is controlled by the polynomial growth order ℓ in $\nabla U(\theta, x)$, see Remark 1. Such an adjustment allows the generalization of the results in [1] to handle superlinear drift coefficients while maintaining the same underlying invariant measure π_*^N. One notes that π_*^N is characterized by the density proportional to $\exp(-\sum_{i=1}^{N} U(\theta, x^i))$ for both systems, as shown in Proposition 1.

2.2. Taming approach

2.2.1 A very brief introduction

For a fixed $T > 0$, consider the SDE given by

$$Y(0) = Y_0, \ dY(t) = b(t, Y(t))dt + \sigma(t, Y(t))dB_t, \ \forall t \in [0, T], \tag{5}$$

where $b(t, y)$ and $\sigma(t, y)$ are assumed to be $\mathcal{B}(\mathbb{R}_+) \times \mathcal{B}(\mathbb{R}^d)-$measurable functions and Y_0 a random variable. The discrete time Markov chain associated with the ULA algorithm is obtained by the Euler-Maruyama discretization scheme of SDE (5) and is defined for all $n \in \mathbb{N}$ by

$$Y(t_0) = Y_0, \ Y(t_{n+1}) = Y(t_n) - \lambda b(t_n, Y(t_n)) + \sqrt{\lambda}\sigma(t_n, Y(t_n))\xi_{n+1}, \ n \geq 0,$$

where $\lambda > 0$ is the stepsize and $(\xi_n)_{n \geq 1}$ is an i.i.d. sequence of standard Gaussian random variables. In the case where the drift coefficient b is superlinear, it is shown in [7] that ULA is unstable, in the sense that any $p-$absolute moment of the algorithm ($p \geq 1$) diverges to infinity in finite time as the stepsize converge to zero. In the literature on SDE approximation, a new class of explicit numerical schemes has been introduced to study the case of non-globally Lipschitz conditions. These schemes modify both the drift and diffusion coefficients such that they grow at most linearly, for example see, [9, 19, 21]. The efficiency of such schemes and their respective properties of $\mathcal{L}^p$ convergence create a strong incentive to extend this technique to sampling and optimization. Typically tamed schemes are given by

$$Y(t_0) = Y_0, \ Y(t_{n+1}) = Y(t_n) - \lambda b_\lambda(t_n, Y(t_n)) + \sqrt{\lambda}\sigma_\lambda(t_n, Y(t_n))\xi_{n+1}, \ n \geq 0,$$

for an appropriate choice of taming functions $b_\lambda : \mathbb{R}^d \mapsto \mathbb{R}^d$ and $\sigma_\lambda : \mathbb{R}^d \mapsto \mathbb{R}^{d \times d}$.

2.2.2 Application in the IPLA framework

We extend the notion of taming to our setting, which involves superlinear drift coefficients, ∇U and constant diffusion coefficients. To achieve this, we introduce a family of taming functions $(h_\lambda)_{\lambda \geq 0}$ with $h_\lambda : \mathbb{R}^{d^\theta} \times \mathbb{R}^{d^x} \to \mathbb{R}^{d^\theta} \times \mathbb{R}^{d^x}$ which are close approximations of ∇U in a manner that is made precise below. We suggest two such taming functions $h_\lambda(v)$. One

uses a uniform taming approach, while the other employs a coordinate-wise method:

$$h_{\lambda,u}(v) = \frac{h(v) - \mu v}{1 + \lambda^{1/2} N^{-p/2} |h(v) - \mu v|} + \mu v, \tag{6}$$

$$h_{\lambda,c}(v) = \left(\frac{h^{(i)}(v) - \mu v^{(i)}}{1 + \lambda^{1/2} |h^{(i)}(v) - \mu v^{(i)}|} + \mu v^{(i)} \right)_{i \in \{1,\dots,d^\theta + d^x\}}, \tag{7}$$

where μ is the strong convexity constant given in A2. In [2], it is experimentally established that the coordinate-wise version outperforms the uniform taming approach. This result is consistent with the observation that uniform taming does not distinguish between the different levels of contribution each coordinate offers to the gradient. However, there is a trade off for using the coordinate-wise approach, as in order to obtain an appropriate dissipativity condition for the tamed function h_λ, additional smoothness on the potential U is required.

We propose two different algorithms for the implementation of interacting particles methods within the tIPLA class, which are determined by the choice of the taming function and the stochastic dynamics. We find that when the uniform taming is used, then the time-scaled dynamics (3)-(4) are more suitable as the new factor $1/N^p$ is essential to guarantee convergence of the algorithms iterates in Wasserstein-2 distance. Hence (3)-(4) paired with (6) leads to the following scheme, for every $n \geq 0$,

$$\theta_0^{\lambda,u} = \theta_0, \ \theta_{n+1}^{\lambda,u} = \theta_n^{\lambda,u} - \frac{\lambda}{N^{p+1}} \sum_{i=1}^{N} h_{\lambda,u}^\theta(\theta_n^{\lambda,u}, X_n^{i,\lambda,u}) + \sqrt{\frac{2\lambda}{N^{p+1}}} \xi_{n+1}^{(0)}, \tag{8}$$

$$X_0^{i,\lambda,u} = x_0^i, \ X_{n+1}^{i,\lambda,u} = X_n^{i,\lambda,u} - \frac{\lambda}{N^p} h_{\lambda,u}^x(\theta_n^{\lambda,u}, X_n^{i,\lambda,u}) + \sqrt{\frac{2\lambda}{N^p}} \xi_{n+1}^{(i)}, \ \text{ for } i = 1, \dots, N. \tag{9}$$

In the case of coordinate-wise taming, the extra smoothness required obviates any modification to the original dynamics. Therefore one uses (1)-(2) with (7) to obtain the second algorithm, for every $n \geq 0$,

$$\theta_0^{\lambda,c} = \theta_0, \ \theta_{n+1}^{\lambda,c} = \theta_n^{\lambda,c} - \frac{\lambda}{N} \sum_{i=1}^{N} h_{\lambda,c}^\theta(\theta_n^{\lambda,c}, X_n^{i,\lambda,c}) + \sqrt{\frac{2\lambda}{N}} \xi_{n+1}^{(0)}, \tag{10}$$

$$X_0^{i,\lambda,c} = x_0^i, X_{n+1}^{i,\lambda,c} = X_n^{i,\lambda,c} - \lambda h_{\lambda,c}^x(\theta_n^{\lambda,c}, X_n^{i,\lambda,c}) + \sqrt{2\lambda} \xi_{n+1}^{(i)}, \ \text{ for } i = 1, \dots, N. \tag{11}$$

It is important to note that, despite the differences in their setups, both algorithms exhibit the same non-asymptotic convergence behavior and dependence on the dimension. These schemes are summarised in the algorithms (tIPLAu) and (tIPLAc) presented below. It is apparent that the newly proposed algorithms, designated as tIPLAu and tIPLAc, incorporate founda-

---

Algorithm 1 Tamed interacting particle Langevin algorithm (tIPLAu)

Require: N,λ,$\pi_{\text{init}} \in \mathcal{P}(\mathbb{R}^{d^\theta}) \times \mathcal{P}((\mathbb{R}^{d^x})^N)$

  Draw $(\theta_0, \{X_0^{i,N}\}_{1 \leq i \leq N})$ from π_{init}

  for n = 0 to $n_T = \lfloor T/\lambda \rfloor$ **do**

  $$\theta_{n+1}^{\lambda,u} = \theta_n^{\lambda,u} - \frac{\lambda}{N^{p+1}} \sum_{i=1}^{N} h_{\lambda,u}^\theta(\theta_n^{\lambda,u}, X_n^{i,\lambda,u}) + \sqrt{\frac{2\lambda}{N^{p+1}}} \xi_{n+1}^{(0)}$$

  $$X_{n+1}^{i,\lambda,u} = X_n^{i,\lambda,u} - \frac{\lambda}{N^p} h_{\lambda,u}^x(\theta_n^{\lambda,u}, X_n^{i,\lambda,u}) + \sqrt{\frac{2\lambda}{N^p}} \xi_{n+1}^{(i)}, \ \forall i \in 1, ..., N$$

  end for

  return θ_{n_T+1}

---

tional elements from [10], which employed deterministic dynamics in the θ component. Furthermore, their design benefits from the approach in [1] which introduced stochastic noise in the gradient flow of θ, enabling explicit calculation of the invariant measure. The key addition of our approach lies in the application of a taming technique via the recalibration of the original drift terms, h_x and h_θ, into tamed counterparts, as described in (6) and (7). This adjustment is crucial in stabilizing the algorithm and deriving non-asymptotic results.

Algorithm 2 Tamed interacting particle Langevin algorithm - coordinate-wise (tIPLAc)

Require: N, $\lambda, \pi_{\text{init}} \in \mathcal{P}(\mathbb{R}^{d^\theta}) \times \mathcal{P}((\mathbb{R}^{d^x})^N)$

  Draw $(\theta_0, \{X_0^{i,N}\}_{1 \leq i \leq N})$ from π_{init}

  for n = 0 to $n_T = \lfloor T/\lambda \rfloor$ **do**

$$\theta_{n+1}^{\lambda,c} = \theta_n^{\lambda,c} - \frac{\lambda}{N} \sum_{i=1}^N h_{\lambda,c}^\theta(\theta_n^{\lambda,c}, X_n^{i,\lambda,c}) + \sqrt{\frac{2\lambda}{N}} \xi_{n+1}^{(0)}$$

$$X_{n+1}^{i,\lambda,c} = X_n^{i,\lambda,c} - \lambda h_{\lambda,c}^x(\theta_n^{\lambda,c}, X_n^{i,\lambda,c}) + \sqrt{2\lambda} \xi_{n+1}^{(i)}, \ \forall i \in 1, ..., N$$

  end for

  return θ_{n_T+1}

2.3. Essential quantities and proof strategy of main result

In this subsection, we lay out the analytical framework and convergence properties of the sequences $(\theta_n^{\lambda,\cdot})_{n \geq 0}$ of either tIPLAu $(\theta_n^{\lambda,u})_{n \geq 0}$ or tIPLAc $(\theta_n^{\lambda,c})_{n \geq 0}$, resulting in their convergence to the maximiser θ^*. A central concept of the technical analysis is the rescaling of the original dynamics of the following form

$$\mathcal{Z}_t^{\cdot,N} = \left(\vartheta_t^{\cdot,N}, N^{-1/2} \mathcal{X}_t^{1,\cdot,N}, \ldots, N^{-1/2} \mathcal{X}_t^{N,\cdot,N} \right), \tag{12}$$

along with the corresponding rescaling that is required for the algorithms and their continuous time interpolations, see more details in Appendix A. This rescaled version of the particle system is essential in bridging the dynamics represented by either (1)-(2) or (3)-(4) with the pairs $\mathcal{V}_t^{i,\cdot,N} := (\vartheta_t^{\cdot,N}, \mathcal{X}_t^{i,\cdot,N})$ generated by the algorithms. The essence of this rescaling lies in its ability to equate the moments of $\mathcal{Z}_t^{\cdot,N}$ with the averaged moments of the pairs $\mathcal{V}_t^{i,\cdot,N}$ across the $\mathcal{L}^2$ norm. This rescaling choice yields the critical property

$$|\mathcal{Z}_t^{\cdot,N}|^2 = \frac{1}{N} \sum_{i=1}^N |\mathcal{V}_t^{i,\cdot,N}|^2.$$

It is imperative to note that this convenient rescaling does not detract from the analysis objectives, as our primary interest is in the convergence of the θ-component. The introduction of particles serves primarily to facilitate sampling, hence their specific scaling does not impact the core convergence analysis.

In the present analysis, the interacting particle systems defined in equations (1)-(2) and (3)-(4) are deliberately constructed to target the same invariant measure π_*^N characterized by the density $Z^{-1} \exp(-\sum_{i=1}^N U(\theta, x^i))$. This aspect is critical as it assigns the number of particles N a role akin to the inverse temperature parameter encountered in simulated annealing algorithms, thereby regulating the concentration of the θ-marginal of the invariant measure π_Θ^N, towards the maximiser θ^*. This behavior, wherein $\pi_\Theta^N \to \delta_{\theta^*}$ as $N \to \infty$, is established in Proposition 2.

Moreover, the convergence rate of the Langevin diffusion to its invariant measure is a standard result, one could consult [5] and the references within. This is further generalized to particle systems in the research conducted by [1]. By integrating their findings into our framework, we arrive at Proposition 3 and 4. Equipped with these guarantees, we obtain the convergence of the underlying dynamics to their invariant measure, and subsequently our analysis focuses to the discretization errors inherent in the proposed schemes.

One expects a dimensional scaling of at least $\mathcal{O}(d^{1/2})$ for the Wasserstein-2 numerical error of Langevin based algorithms, for example [6]. In the context of the E optimization where N particles are generated, the true dimensionality of the problem is $d^\theta + N \cdot d^x$, which can be naively interpreted as the need to increase the number of iterations to achieve convergence as N increases. However, this dependency on N is effectively mitigated by making use of the implied symmetry within the dynamics of the interacting particle system by considering only the θ-marginal in the analysis. Such an approach enables us to attain the classical Euler-Maruyama convergence rate for the numerical solutions of stochastic differential equations, as demonstrated in Proposition 5 and 6.

In closing, one decomposes the global $\mathcal{L}^2$-error, denoted as $\left(\mathbb{E} \left[|\theta^* - \theta_n^{\lambda,\cdot}|^2 \right] \right)^{1/2}$, into three distinct components

$$\mathbb{E}^{1/2} \left[|\theta^* - \theta_n^{\lambda,\cdot}|^2 \right] \leq W_2(\delta_{\theta^*}, \pi_\Theta^N) + W_2(\pi_\Theta^N, \mathcal{L}(\vartheta_{n\lambda}^{\cdot,N})) + W_2(\mathcal{L}(\vartheta_{n\lambda}^{\cdot,N}), \mathcal{L}(\theta_n^{\lambda,\cdot})),$$

where $\theta_n^{\lambda,\cdot}$ stands for the iterates of either of the algorithms, tIPLAu and tIPLAc alike. Moreover, $W_2(\delta_{\theta^*}, \pi_\Theta^N)$ quantifies the deviation of the invariant measure from the maximiser, $W_2(\pi_\Theta^N, \mathcal{L}(\vartheta_{n\lambda}^{\cdot,N}))$ captures the discrepancy between the law of the dynamics and their invariant measure, and $W_2(\mathcal{L}(\vartheta_{n\lambda}^{\cdot,N}), \mathcal{L}(\theta_n^{\lambda,\cdot}))$ encompasses the error induced by the discretization scheme, namely, the divergence between the law of the iterations of our algorithms and their continuous counterparts.

3. Main Assumptions and Results

In this section, we provide the assumptions regarding the potential $U(u)$ and its gradient $h(u) = \nabla U(u)$ that define the framework, under which the main results for the non-asymptotic behavior of the newly proposed algorithms, tIPLAu and tIPLAc, are derived.

Let $d := d^\theta + d^x$, and U be a $C^1\left(\mathbb{R}^d\right)$ function. We require that $h(v) := h(\theta, x)$ is a locally Lipschitz function in θ and x.

A1. There exist $L > 0$ and $\ell > 0$ such that

$$|h(v) - h(v')| \leq L\left(1 + |v|^\ell + |v'|^\ell\right)|v - v'|, \; \forall v, v' \in \mathbb{R}^d.$$

Additionally we require U to be $\mu-$strongly convex.

A2. There exists $\mu > 0$ such that

$$\langle v - v', h(v) - h(v')\rangle \geq \mu|v - v'|^2, \; \forall v, v' \in \mathbb{R}^d.$$

Assumption A1 is a significant relaxation of the global Lipschitz condition which is widely used in the literature. It allows the gradient h to grow polynomially fast at infinity. Assumption A2 guarantees that U has a unique minimiser and also can be seen as a monotonicity type condition, which is satisfied by the drift coefficients of the corresponding SDEs.

We also present a growth condition for the gradient ∇U, which is essential in the case where the coordinate-wise taming function (7) is used.

A3-i. For each i in $\{1, \ldots, d\}$ there exists $\mu > 0$ such that

$$h^{(i)}(v)v^{(i)} \geq \frac{\mu}{2}|v^{(i)}|^2 - \frac{1}{2\mu}|h_0^{(i)}|^2.$$

In the special case where $d^x = d^\theta$ A3-i can be slightly relaxed to the following form.

A3-ii. Let $d^x = d^\theta = d$, then for each i in $\{1, \ldots, d\}$ there exist $\mu > 0$, $\rho \geq 0$ satisfying the constrain $4\rho < \mu$, such that

$$h^{\theta_i}(v)\theta_i \geq \frac{\mu}{2}|\theta_i|^2 - \rho|x_i|^2 - \frac{1}{2\mu}|h_0^{\theta_i}|^2 \text{ and } h^{x_i}(v)x_i \geq \frac{\mu}{2}|x_i|^2 - \rho|\theta_i|^2 - \frac{1}{2\mu}|h_0^{x_i}|^2.$$

Assumption A3 introduces a coordinate-wise dissipativity type condition, which plays a critical role in establishing moment bounds for tIPLAc. Its necessity stems from the fact that Remark 2 does not guarantee that the taming function (7) preserves the dissipativity of the drift $h(v) = \nabla U(v)$.

We emphasize that Assumption A3 is primarily influenced by the structural properties of the taming function. This assumption could potentially be relaxed to require only argument-wise dissipativity, as described by the following mathematical formulation:

$$h^\theta(v)\theta \geq \frac{\mu}{2}|\theta|^2 - \rho|x|^2 - \frac{1}{2\mu}|h_0^\theta|^2 \text{ and } h^x(v)x \geq \frac{\mu}{2}|x|^2 - \rho|\theta|^2 - \frac{1}{2\mu}|h_0^x|^2.$$

This relaxation is applicable when considering the taming function

$$h_{\lambda,a}^w(v) = \frac{h^w(v) - \mu w}{1 + \lambda^{1/2}|h^w(v) - \mu w|} + \mu w \text{ for } w = \theta, \; x,$$

which represents an intermediate scenario bridging the uniform and coordinate-wise approaches. Such consideration leads to an alternative formulation of the tIPLA algorithm. However, this variant of the tIPLA algorithm is not of primary interest

6

due to its inherent limitations. In particular, it fails to adequately discriminate between the components of the gradient that exhibit explosive behavior and those that do not. While simultaneously, this alternative formulation imposes more stringent requirements than those necessitated by the uniform case. Lastly, we proceed with an assumption on starting values for the tIPLA algorithms.

A4. The initial condition $z_0^N = \left(\theta_0, N^{-1/2}x_0^1, \ldots, N^{-1/2}x_0^N\right)$ is such that

$$\mathbb{E}\left[|z_0^N|^{2p_0}\right] < \infty,$$

where $p_0 = 2(\ell + 1)$ and ℓ is the order of polynomial growth of $\nabla U(\theta, x)$ as given in Assumption A1.

Remark 1. *Assumption A1 imposes a restriction on the growth of $h(v) = \nabla U(v)$, that is*

$$h(v) \leq K(1 + |v|^{\ell+1}), \ \forall v \in \mathbb{R}^d, \tag{13}$$

where $K = 2L + |h_0|$.

Remark 2. *From Assumption A2 (strong convexity) one deduces that it implies dissipativity*

$$\langle v, h(v) \rangle \geq \frac{\mu}{2}|v|^2 - b, \ \forall v \in \mathbb{R}^d, \tag{14}$$

where $b = \dfrac{1}{2\mu}|h_0|^2$.

3.1. Taming functions and inherited properties

In this subsection, we establish important properties of the proposed taming functions.

Property 1. For all $\lambda > 0$ and $v \in \mathbb{R}^d$, it holds that

$$|h_{\lambda,u}(v)| \leq \mu|v| + \lambda^{-1/2}N^{p/2},$$
$$|h_{\lambda,c}(v)| \leq \mu|v| + (d^\theta + d^x)\lambda^{-1/2}.$$

The gradient h is sufficiently close to its tamed counterparts $h_{\lambda,u}, h_{\lambda,c}$ for $\lambda > 0$ small enough.

Property 2. For all $\lambda > 0$ and $v \in \mathbb{R}^d$, it holds that

$$|h_{\lambda,u}(v) - h(v)| \leq C_1\lambda^{1/2}N^{-p/2}(1 + |v|^{2(\ell+1)}),$$
$$|h_{\lambda,c}(v) - h(v)| \leq C_1\lambda^{1/2}(1 + |v|^{2(\ell+1)}).$$

where $C_1 = 2^{2(\ell+3/2)}\max\{K^2, \mu^2\}$.

The taming functions $h_{\lambda,u}, h_{\lambda,c}$ inherit the dissipativity condition established in Remark 2.

Property 3. For all $\lambda > 0$ and $v \in \mathbb{R}^d$, it holds that

$$\langle v, h_{\lambda,u}(v) \rangle \geq \frac{\mu}{2}|v|^2 - b, \text{ and } \langle v, h_{\lambda,c}(v) \rangle \geq \frac{\mu}{2}|v|^2 - b,$$

where b is given in Remark 2. We note that in the case of the coordinate-wise taming function to be dissipative, we require the additional smoothness provided by Assumption A3.

In the context of a unified framework for Langevin based algorithm under superlinear gradients, as established in [11], the properties 1-3 therein are recovered for $\delta = 2$ and $\gamma = 1/2$ which aligns with THεO POULA [13] and TUSLA [14] algorithms.

3.2. Preliminary Theoretical Statements

Both interacting SDEs defined by (1)-(2) and (3)-(4) respectively, admit a strong solution under A1-A2. This follows since each of the drift coefficients of the SDEs are locally Lipschitz functions and dissipative, in view of Remark 2, whereas the diffusion coefficients are constant. One could consult [15] (Theorem 2.3.5) for more details. Moreover, both systems exhibit the same invariant measure, as one would expect from a standard Langevin diffusion.

Proposition 1. *Let A1 and A2 hold. Then, the measure π_*^N characterized by the density $Z^{-1} \exp\left(-\sum_{i=1}^{N} U(\theta, x^i)\right)$, with Z being the normalizing constant, is the invariant measure for both interacting particles systems* (1)-(2) *and* (3)-(4).

Proof. The existence of an invariant measure is established by invoking the Krylov-Bogoliubov Theorem, as presented in Theorem 7.1 [17]. This theorem guarantees the existence of an invariant measure for a Markov process generated by its corresponding semigroup, given a tightness condition on the sequence of probability measured associated with that process. According to Da Prato's Proposition 7.10 [17], the establishment of uniform moment bounds, provided in Lemma 1, serves as a sufficient condition for tightness. The uniqueness of the invariant measure follows from Theorem 7.16 (ii) [17], as Hypothesis 7.13 is satisfied due to A1 and A2. Furthermore, one can verify that π_*^N is indeed the invariant measure in discussion by checking that for any $\phi \in C_c^\infty(\mathbb{R}^d)$ one has $\int_{\mathbb{R}^d} \mathcal{L}\phi \, d\pi_*^N = 0$, where $\mathcal{L}$ is the infinitesimal generator of the process. See the proof of Proposition 2 in [1] for more details. $\square$

Henceforth we consider exclusively the θ−marginal of the invariant measure. We show that indeed the number of particles N, plays the role of the inverse temperature parameter in temperature annealing algorithms in the sense that it provides control over the marginal's concentration around the maximiser.

Proposition 2. *Let π_Θ^N denote the θ−marginal of the invariant measure and θ^* be the maximiser of $k(\theta)$. Then, under A2, for any $N \in \mathbb{N}$, the following bound holds*

$$W_2(\pi_\Theta^N, \delta_{\theta^*}) \leq \sqrt{\frac{d^\theta}{\mu N}}.$$

Proof. Follows from Proposition 3 in [1], with the observation that strong log-concavity is preserved under marginalization, see Theorem 3.8 in [22]. $\square$

Next, we need the following ergodicity result to obtain the convergence of the law of the IPS to the invariant measure.

Proposition 3. *Let A1, A2 and A4 hold and consider the first component, namely $\vartheta_t^{u,N}$, of the continuous dynamics* (3)-(4). *Then, for any $N \in \mathbb{N}$,*

$$W_2\left(\mathcal{L}(\vartheta_t^{u,N}), \pi_\Theta^N\right) \leq \exp\left(-\mu t/N^p\right) \left(\left(\mathbb{E}\left[|z_0^N - z^*|^2\right]\right)^{1/2} + \left(\frac{d^x N + d^\theta}{\mu N^{p+1}}\right)^{1/2}\right).$$

Proof. The proof is postponed to Appendix A. $\square$

Proposition 4. *Let A1-A4 hold and consider the first component, namely $\vartheta_t^{c,N}$, of the continuous dynamics* (1)-(2). *Then, for any $N \in \mathbb{N}$,*

$$W_2\left(\mathcal{L}(\vartheta_t^{c,N}), \pi_\Theta^N\right) \leq \exp\left(-\mu t\right) \left(\left(\mathbb{E}\left[|z_0^N - z^*|^2\right]\right)^{1/2} + \left(\frac{d^x N + d^\theta}{\mu N}\right)^{1/2}\right).$$

Proof. Follows closely from the proof of Proposition 4 in [1]. One can easily verify that although the corresponding A1 in [1] invokes global Lipschitz continuity rather than local, its need there is solely to ensure the existence of a unique global solution, a fact already established at the start of Subsection 3.2. $\square$

3.3. Discretisation Error Estimates

We consider the continuous interpolation of the tIPLAu algorithm, which approximates the time-scaled version of the continuous dynamics (3)-(4), denoted by $\overline{Z}_t^{\lambda,u}$. Notably, the law of the discretized process and its interpolation coincide at

grid-points, i.e. $\mathcal{L}(Z_n^{\lambda,u}) = \mathcal{L}(\overline{Z}_n^{\lambda,u})$. The continuous-time interpolation of tIPLAu is defined as follows

$$\overline{\theta}_0^{\lambda,u} = \theta_0, \ d\overline{\theta}_t^{\lambda,u} = -\frac{\lambda}{N^{p+1}} \sum_{i=1}^N h_{\lambda,u}^\theta(\theta_{\lfloor t \rfloor}^{\lambda,u}, X_{\lfloor t \rfloor}^{i,\lambda,u})dt + \sqrt{\frac{2\lambda}{N^{p+1}}} dB_t^{0,\lambda}, \tag{15}$$

$$\overline{X}_0^{i,\lambda,u} = x_0^i, \ d\overline{X}_t^{i,\lambda,u} = -\frac{\lambda}{N^p} h_{\lambda,u}^x(\theta_{\lfloor t \rfloor}^{\lambda,u}, X_{\lfloor t \rfloor}^{i,\lambda,u})dt + \sqrt{\frac{2\lambda}{N^p}} dB_t^{i,\lambda}, \ \text{for } i = 1,..,N, \tag{16}$$

whereas the time-changed SDEs (3)-(4) are given by

$$d\vartheta_{\lambda t}^{u,N} = -\frac{\lambda}{N^{p+1}} \sum_{i=1}^N h^\theta(\vartheta_{\lambda t}^{u,N}, \mathcal{X}_{\lambda t}^{i,u,N})dt + \sqrt{\frac{2\lambda}{N^{p+1}}} dB_t^{0,\lambda}, \tag{17}$$

$$d\mathcal{X}_{\lambda t}^{i,u,N} = -\frac{\lambda}{N^p} h^x(\vartheta_{\lambda t}^{u,N}, \mathcal{X}_{\lambda t}^{i,u,N})dt + \sqrt{\frac{2\lambda}{N^p}} dB_t^{i,\lambda}, \ \text{for } i = 1,..,N, \tag{18}$$

where $B_t^{j,\lambda} := B_{\lambda t}^j / \sqrt{\lambda}$, $t \geq 0$, $0 \leq j \leq N$, represents a Brownian motion under its completed natural filtration $\mathcal{F}_t^{j,\lambda} := \mathcal{F}_{\lambda t}^j$.

Proposition 5. *Let A1, A2 and A4 hold and consider the process $\vartheta_{\lambda t}^{u,N}$ as given by the dynamics (17)-(18). Then, for every $\lambda_0 < N^{2\ell+1}/(4\mu)$, there exists a constant $C > 0$ independent of N, n, λ, d^x and d^θ such that for any $\lambda \in (0, \lambda_0)$,*

$$\left(\mathbb{E}\left[|\overline{\theta}_n^{\lambda,u} - \vartheta_{n\lambda}^{u,N}|^2 \right] \right)^{1/2} \leq \lambda^{1/2} C_{|z_0|,\mu,b,\ell,L}(1 + d^\theta/N + d^x)^{\ell+1},$$

for all $n \in \mathbb{N}$.

Proof. The proof is postponed to Appendix A. $\square$

Similarly, we define the corresponding auxiliary processes for the iterates of tIPLAc. The continuous-time interpolation is given by

$$\overline{\theta}_0^{\lambda,c} = \theta_0, \ d\overline{\theta}_t^{\lambda,c} = -\frac{\lambda}{N} \sum_{i=1}^N h_{\lambda,c}^\theta(\theta_{\lfloor t \rfloor}^{\lambda,c}, X_{\lfloor t \rfloor}^{i,\lambda,c})dt + \sqrt{\frac{2\lambda}{N}} dB_t^{0,\lambda}, \tag{19}$$

$$\overline{X}_0^{i,\lambda,c} = x_0^i, \ d\overline{X}_t^{i,\lambda,c} = -\lambda h_{\lambda,c}^x(\theta_{\lfloor t \rfloor}^{\lambda,c}, X_{\lfloor t \rfloor}^{i,\lambda,c})dt + \sqrt{2\lambda} dB_t^{i,\lambda}, \ \text{for } i = 1,\ldots,N, \tag{20}$$

and the time changed SDEs of (1)-(2) are

$$d\vartheta_{\lambda t}^{c,N} = -\frac{\lambda}{N} \sum_{i=1}^N h^\theta(\vartheta_{\lambda t}^{c,N}, \mathcal{X}_{\lambda t}^{i,c,N})dt + \sqrt{\frac{2\lambda}{N}} dB_t^{0,\lambda}, \tag{21}$$

$$d\mathcal{X}_{\lambda t}^{i,c,N} = -\lambda h^x(\vartheta_{\lambda t}^{c,N}, \mathcal{X}_{\lambda t}^{i,c,N})dt + \sqrt{2\lambda} dB_t^{i,\lambda}, \ \text{for } i = 1,\ldots,N, \tag{22}$$

where the Brownian motions are defined as in (17)-(18).

Proposition 6. *Let A1, A2, A3-i and A4 hold and consider the process $\vartheta_{\lambda t}^{c,N}$ as given by the dynamics (21)-(22). Then, for every $\lambda_0 < 1/(4\mu)$, there exists a constant $C > 0$ independent of N, n, λ, d^x and d^θ such that for any $\lambda \in (0, \lambda_0)$,*

$$\left(\mathbb{E}\left[|\overline{\theta}_n^{\lambda,c} - \vartheta_{n\lambda}^{c,N}|^2 \right] \right)^{1/2} \leq \lambda^{1/2} C_{|z_0|,\mu,b,\ell,L}(1 + d^\theta + d^x)^{\ell+1},$$

for all $n \in \mathbb{N}$. If assumption A3-ii is used in place of A3-i, the proposition holds under the updated timestep constraint, $\lambda_0 < 1/(8\mu)$.

Proof. The proof is postponed to Appendix A. $\square$

3.4. Main results and global error

Theorem 1. *Consider the iterates $(\theta_n^{\lambda,u})_{n\geq 0}$ as given in tIPLAu, and let A1, A2 and A4 hold. Then, for every $\lambda_0 < N^{2\ell+1}/(4\mu)$, there exists a constant $C > 0$ independent of N, n, λ, d^x and d^θ such that, for any $\lambda \in (0, \lambda_0)$,*

$$
\left(\mathbb{E}\left[|\theta^* - \theta_n^{\lambda,u}|^2\right]\right)^{1/2} \leq \sqrt{\frac{d^\theta}{\mu N}} + \lambda^{1/2} C_{|z_0|,\mu,b,\ell,L}(1 + d^\theta/N + d^x)^{\ell+1}
$$

$$
+ \exp\left(-\mu n\lambda/N^{2\ell+1}\right)\left(\left(\mathbb{E}\left[|z_0^N - z^*|^2\right]\right)^{1/2} + \left(\frac{d^x N + d^\theta}{\mu N^{2(\ell+1)}}\right)^{1/2}\right),
$$

for all $n \in \mathbb{N}$.

Proof. Combining the results from Propositions 2, 3 and 5, we decompose the expectation into three terms: one describing the concentration of π_Θ^N around θ^*, one describing the convergence of the IPS to its invariant measure, and lastly one describing the error induced by the time discretisation

$$
\left(\mathbb{E}\left[|\theta^* - \theta_n^{\lambda,u}|^2\right]\right)^{1/2} = W_2(\delta_{\theta^*}, \mathcal{L}(\theta_n^{\lambda,u}))
$$

$$
\leq W_2(\delta_{\theta^*}, \pi_\Theta^N) + W_2(\pi_\Theta^N, \mathcal{L}(\vartheta_{n\lambda}^{u,N})) + W_2(\mathcal{L}(\vartheta_{n\lambda}^{u,N}), \mathcal{L}(\theta_n^{\lambda,u})). \tag{23}
$$

Substituting the bounds from the aforementioned Propositions yields the final inequality. $\square$

Theorem 2. *Consider the iterates $(\theta_n^{\lambda,c})_{n\geq 0}$ as given in tIPLAc, and let A1, A2, A3-i and A4 hold. Then, for every $\lambda_0 < 1/(4\mu)$, there exists a constant $C > 0$ independent of N, n, λ, d^x and d^θ such that, for any $\lambda \in (0, \lambda_0)$,*

$$
\left(\mathbb{E}\left[|\theta^* - \theta_n^{\lambda,c}|^2\right]\right)^{1/2} \leq \sqrt{\frac{d^\theta}{\mu N}} + \exp\left(-\mu n\lambda\right)\left(\left(\mathbb{E}\left[|z_0^N - z^*|^2\right]\right)^{1/2} + \left(\frac{d^x N + d^\theta}{\mu N}\right)^{1/2}\right)
$$

$$
+ \lambda^{1/2} C_{|z_0|,\mu,b,\ell,L}(1 + d^\theta + d^x)^{\ell+1},
$$

for all $n \in \mathbb{N}$. If assumption A3-ii is used in place of A3-i, the theorem holds under the updated timestep constraint, $\lambda_0 < 1/(8\mu)$.

Proof. To prove the above inequality, we replace the bound for $W_2(\pi_\Theta^N, \mathcal{L}(\vartheta_{n\lambda}^{\cdot,N}))$ in (23) given by Proposition 3 with the bound given by Proposition 4 and respectively the bound for $W_2(\mathcal{L}(\vartheta_{n\lambda}^{\cdot,N}), \mathcal{L}(\theta_n^{\lambda,\cdot}))$ given by Proposition 5 with the bound given by Proposition 6. $\square$

It is observed that the dependence on the dimension d^θ of the parameter space in Theorem 2 exhibits a slight deterioration compared to that presented in Theorem 1 and in Theorem 1 of [1]. Notably, the factor $1/N$ is absent in the third term of Theorem 2. This is interpreted as a compromise entailed by the use of the coordinate-wise taming function (7), which requires the direct derivation of moment bounds for the pairs $\mathcal{V}_t^{i,\cdot,N}$, as discussed in Subsection 2.3. Consequently, this approach results in the loss of the symmetrical structure that is attained by considering the N-particle system $\mathcal{X}_t^{i,\cdot,N}$ in its entirety.

Corollary 1. *Consider the iterates of tIPLAu, $(\theta_n^{\lambda,u})_{n\geq 0}$ and let $\epsilon > 0$. Then, for $N \geq 9d^\theta/(\mu\epsilon^2)$ and $\lambda \leq \min\{\lambda_0, \epsilon^2/(9C_0(1 + \mu\epsilon^2/9 + d^x)^{2\ell+2})\}$, one needs $n \geq (9/\mu)^{2\ell+2}((d^\theta)^{2\ell+1}/\epsilon^{4\ell+4})C_0^2(1 + \mu\epsilon^2/9 + d^x)^{2\ell+2}\log(3(R_0 + (d^x\epsilon^{4\ell+2}\mu^{2\ell}/(9d^\theta)^{2\ell+1} + 9\epsilon^{4\ell+4}\mu^{2\ell+1}/(9d^\theta)^{2\ell+1})^{1/2})/\epsilon)$ iterations to achieve*

$$
W_2(\pi_\beta, \mathcal{L}(\theta_n^{\lambda,u})) \leq \epsilon,
$$

where $R_0 = \left(\mathbb{E}\left[|z_0^N - z^*|^2\right]\right)^{1/2}$. Overall one has to choose $N = \mathcal{O}(1/\epsilon^2)$, $\lambda = \mathcal{O}(\epsilon^2)$ and $n = \mathcal{O}(\epsilon^{-4(\ell+1)}\log(1/\epsilon))$. The corresponding dimensional dependencies are as follows, $N = \mathcal{O}(d^\theta)$, $\lambda = \mathcal{O}((d^x)^{-2(\ell+1)})$ and $n = \mathcal{O}((d^\theta d^x)^{2\ell+1}\log((d^x)^{1/2}/(d^\theta)^{\ell+1/2}))$.

Corollary 2. *Consider the iterates of tIPLAc, $(\theta_n^{\lambda,c})_{n\geq 0}$ and let $\epsilon > 0$. Then, for $N \geq 9d^\theta/(\mu\epsilon^2)$ and $\lambda \leq \min\{\lambda_0, \epsilon^2/(9C_0(1 + d^\theta + d^x)^{2\ell+2})\}$, one needs $n \geq (9/(\mu\epsilon^2))C_0^2(1 + d^\theta + d^x)^{2\ell+2}\log(3(R_0 + (d^x/\mu + \epsilon^2/9)^{1/2})/\epsilon)$ iterations to achieve*

$$W_2(\pi_\beta, \mathcal{L}(\theta_n^{\lambda,c})) \leq \epsilon,$$

where $R_0 = \left(\mathbb{E}\left[|z_0^N - z^*|^2\right]\right)^{1/2}$. Overall one has to choose $N = \mathcal{O}(1/\epsilon^2)$, $\lambda = \mathcal{O}(\epsilon^2)$ and $n = \mathcal{O}(\epsilon^{-2}\log(1/\epsilon))$. The corresponding dimensional dependencies are as follows, $N = \mathcal{O}(d^\theta)$, $\lambda = \mathcal{O}((d^x + d^\theta)^{-2(\ell+1)})$ and $n = \mathcal{O}((d^x + d^\theta)^{2\ell+2}\log(d^x))$.

3.5. Summary of contributions and discussion

This article aims to expand the interacting particle-based gradient decent methods for marginal likelihood optimization beyond linear growth for the corresponding gradients. The contributions of our work can be summarized as follows:

- We provide rigorous results for the treatment of interacting particle systems with superlinear drift coefficients.

- For stepsize λ, we achieve $\lambda^{1/2}$ rate in W_2 distance for both tIPLA-based algorithms to the target measure. To the best of our knowledge, this is the first result under such relaxed assumptions.

- We provide explicit dimensional dependencies in our constants, exhibiting polynomial scaling determined by the polynomial growth of the drift coefficients.

Table 1: Comparison of algorithmic complexity across existing literature

	W_2	CONVEXITY	GROWTH
PGD [10]	$\mathcal{O}\left(\epsilon^{-2}\log(1/\epsilon)\right)$	strongly convex	linear
IPLA [1]	$\mathcal{O}\left(\epsilon^{-2}\log(1/\epsilon)\right)$	strongly convex	linear
tIPLAu	$\mathcal{O}\left(\epsilon^{-4(\ell+1)}\log(1/\epsilon)\right)$	strongly convex	polynomial
tIPlAc	$\mathcal{O}\left(\epsilon^{-2}\log(1/\epsilon)\right)$	strongly convex	polynomial

In Table 1 we compare the performance of our algorithms with other IPS gradient decent variants that do not accommodate superlinear settings. We can see that our result for tIPLAu does not achieve the same complexity as its competitors. This is a natural drawback since we are not limited to the globally Lipschitz continuous case for gradients. However, tIPLAc compares favorably, attaining the same complexity as its competitors by allowing superlinear growth while imposing element-wise dissipativity. We highlight that, SOUL [3] is omitted from Table 1 due to different methodologies being employed: it operates sequentially with a single particle for latent posterior sampling, whereas our approach uses N particles and is parallelizable; additionally, SOUL [3] uses a decreasing stepsize approach to characterize almost sure convergence while in this paper a constant stepsize is used to obtain $\mathcal{L}^2$ convergence estimates. Furthermore, we do not include classical MCMC methods in our numerical comparison, as our focus is on gradient-based approaches.

The tIPLA algorithm class generalizes IPLA [1] by allowing a wider class of potentials with superlinear gradients to be considered. One observes that elements of the original proof structure appear in this work, while new techniques address challenges unique to the new setting. In particular, the main challenge in the superlinear context is designing a taming function that satisfies critical properties, specifically, linear growth (Property 1) and inheriting the dissipativity of the original gradient (Property 3). Both properties are essential for establishing higher moment bounds, which is key to our proof strategy.

The steps, to demonstrate both the concentration of the θ-marginal of the invariant measure around the maximizer and the contraction of the IPS, parallel the corresponding proofs for IPLA. The primary differences lie in the technical modifications required to overcome the challenges introduced by aforementioned taming functions. To control the discretization error, we introduce the time-scaling factor N^p in equations (3)–(4). In the uniform taming case, this factor enables us to bound the sum of the higher moments of the paired processes $V_t^{i,\lambda,u}$ by the corresponding rescaled process

$Z_t^{\lambda,u}$ (see equation (44)). In contrast, for coordinate-wise taming, such a modification is unnecessary because higher moment bounds for $V_t^{i,\lambda,c}$ can be derived directly due to the smoothness Assumption 3.

4. Numerical Experiments

This section presents the performance evaluation of tIPLAu and tIPLAc by applying them on a set optimization problems. We also include competitive algorithms such as IPLA [1], PGD [10], and SOUL [3] for comparison. The results highlight the ability of our methods to operate under more relaxed assumptions in the superlinear setting, where other algorithms face limitations.

4.1. Bayesian logistic regression with thin tailed priors

Let $y \in \mathbb{R}^{d^y}$ represent the observable variable, which consists of a set of binary responses $\{y_j\}_{j=1}^{d^y} \in \{0,1\}$, and $x \in \mathbb{R}^{d^x}$, $\theta \in \mathbb{R}^{d^\theta}$ the latent variable and the target parameter to be estimated respectively with $d^x = d^\theta$. We consider the Bayesian logistic regression model where the conditional distribution of y given x is defined as

$$p_\theta(y|x) = \prod_{j=1}^{d^y} s(u_j^T x)^{y_j} (1 - s(u_j^T x))^{1-y_j},$$

where $s(t) = \exp(t)/(1 + \exp(t))$ and $\{u_j\}_{j=1}^{d^y} \in \mathbb{R}^{d^x}$ are fixed covariates corresponding to the observable data. Instead of employing a standard Gaussian prior, we model the prior distribution using a generalized Gaussian with thinner tails, expressed as

$$p_\theta(x) = C_1 \exp\left(-\frac{1}{\sigma^4} \sum_{j=1}^{d^x} |x_j - \theta_j|^4 - \frac{1}{\sigma^2} \sum_{j=1}^{d^x} |x_j - \theta_j|^2\right). \tag{24}$$

Thus, the marginal likelihood can be written as

$$p_\theta(y) = \int_{\mathbb{R}^{d^x}} p_\theta(y|x) p_\theta(x) dx = \int_{\mathbb{R}^{d^x}} \exp\left(-U_y(\theta, x)\right) dx,$$

where

$$U_y(\theta, x) = C_2 + \frac{1}{\sigma^4} \sum_{j=1}^{d^x} |x_j - \theta_j|^4 + \frac{1}{\sigma^2} \sum_{j=1}^{d^x} |x_j - \theta_j|^2 - \sum_{j=1}^{d^y} \left(y_j \ln(s(u_j^T x)) + (1 - y_j) \ln(s(-u_j^T x))\right).$$

Here the constant C_2 consists of terms not dependent on either x or θ. Assuming that σ^2 is a fixed, known constant, we verify that the potential $U_y(\theta, x)$ satisfies our main assumptions A1 and A2, in view of Remark 3 and Remark 4.

Experiment details. Following a similar setup with [1], we choose $d^x = d^\theta = 3$ and $d^y = 900$ and set $\theta^* = [2, 5, -1]$ to test the convergence of the parameter. We run tIPLAu for $M = 200\text{k}$ steps, using $N = 10, 100, 1000$ particles and a stepsize of $\lambda = 0.0001$. We set $\sigma^2 = 0.1$ and simulate the covariates v_j from the uniform distribution $U(-1, 1)$ over $\mathbb{R}^{d^x}$. The parameter θ_0 is initialized from a deterministic value $[100, -100, 0]$, while the particles $X_0^{i,N}$ are drawn from the generalized Gaussian (24) with mean θ_0 and variance $\sigma^2 = 0.1$.

As shown in Fig. 1, we observe that as N increases, the convergence time to the true value decreases, with a pronounced effect when N exceeds 1000. However, as seen in Fig.2, which displays the last 10,000 iterations of each simulation, there is no significant impact on the residual bias of the algorithm once convergence has been achieved. This suggests that the source of the bias in this case comes from the discretization error, which remains largely the same with respect to N and is primarily dependent on the problem-specific constants and the chosen stepsize. Moreover, it is evident that as N increases, the Markov chains generated by the algorithm become less noisy, as one would expect from Langevin based algorithms when the inverse temperature parameter is increased.
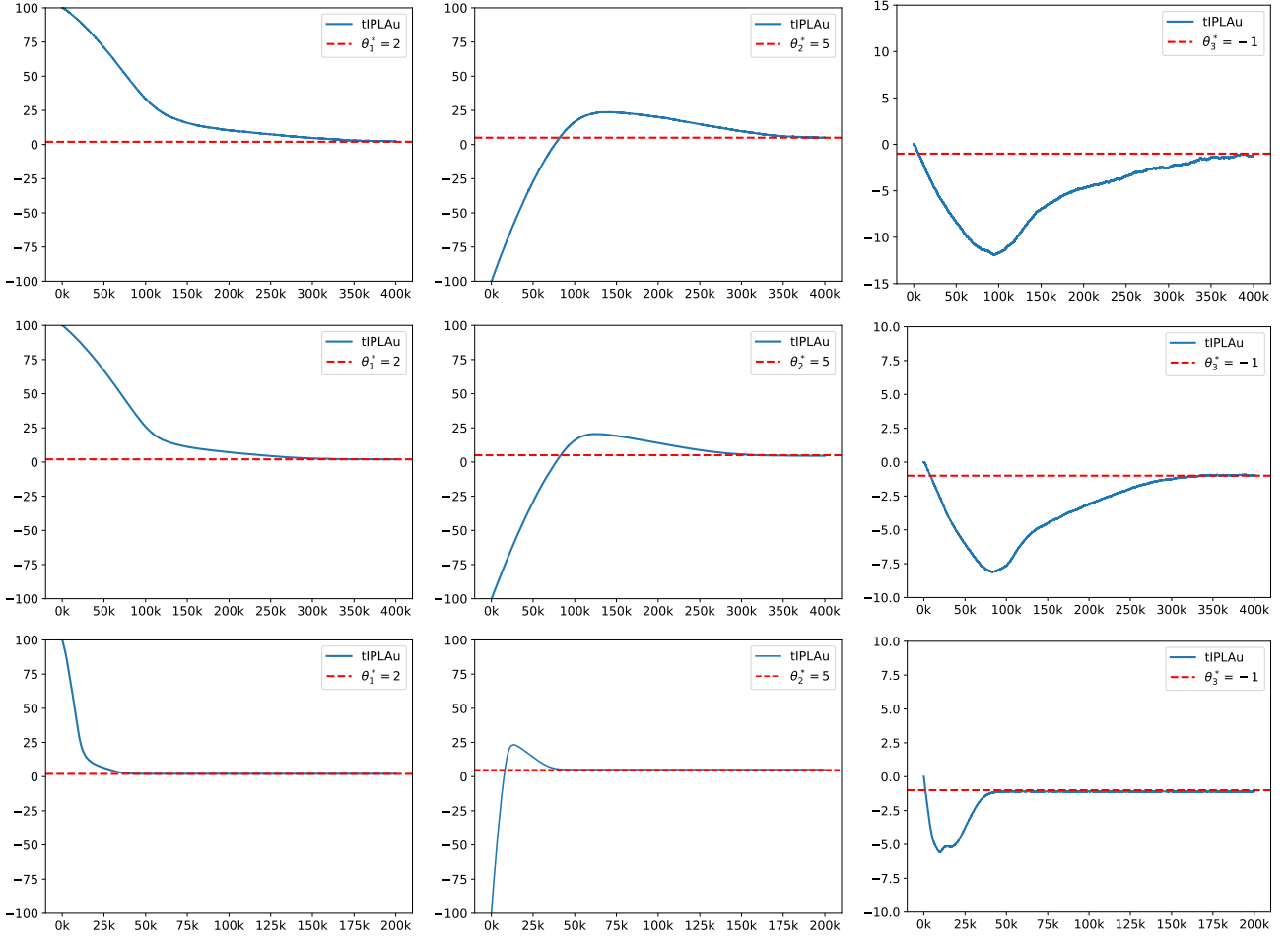
Figure 1: The performance of tIPLAu on the synthetic logistic regression problem. Each column corresponds to one of the components of the true parameter $\theta^* = [2, 5, -1]$. The top row corresponds to $N = 10$, the middle row to $N = 100$, and the bottom row to $N = 1000$

4.2. Higher order - Toy example

To demonstrate the applicability and versatility of our framework we consider the following highly superlinear toy example where an intractable integral has to be maximized with respect to the parameter θ. Let for any $m \geq 1$ the objective function to be defined as

$$k(\theta) = \int_{\mathbb{R}^{d^x}} \exp\left(-|x|^{4m} - (|x|^{2m} + 1)(|\theta|^{2m} + 1) - |\theta|^{4m}\right.$$
$$\left. -|x|^4 - (|x|^2 + 1)(|\theta|^2 + 1) - |\theta|^4\right) dx.$$

In this case the maximiser $\theta^* = \operatorname{argmax}\{k(\theta)\}$ can easily be derived to be $\theta^* = 0$ which helps us verify our experimental foundings. One simply observes that

$$\exp\left(-|x|^{4m} - (|x|^{2m} + 1)(|\theta|^{2m} + 1) - |\theta|^{4m} - |x|^4 - (|x|^2 + 1)(|\theta|^2 + 1) - |\theta|^4\right)$$
$$\leq \exp\left((-|x|^{4m} - (|x|^{2m} + 1) - |x|^4 - (|x|^2 + 1)\right),$$

with the equality holding if and only if $\theta = 0$. Hence by taking the definite integral with respect to dx over the whole $\mathbb{R}^{d^x}$ we have

$$k(\theta) \leq \int_{\mathbb{R}^{d^x}} \exp\left(-|x|^{4m} - (|x|^{2m} + 1) - |x|^4 - (|x|^2 + 1)\right) dx := k(\theta^*).$$
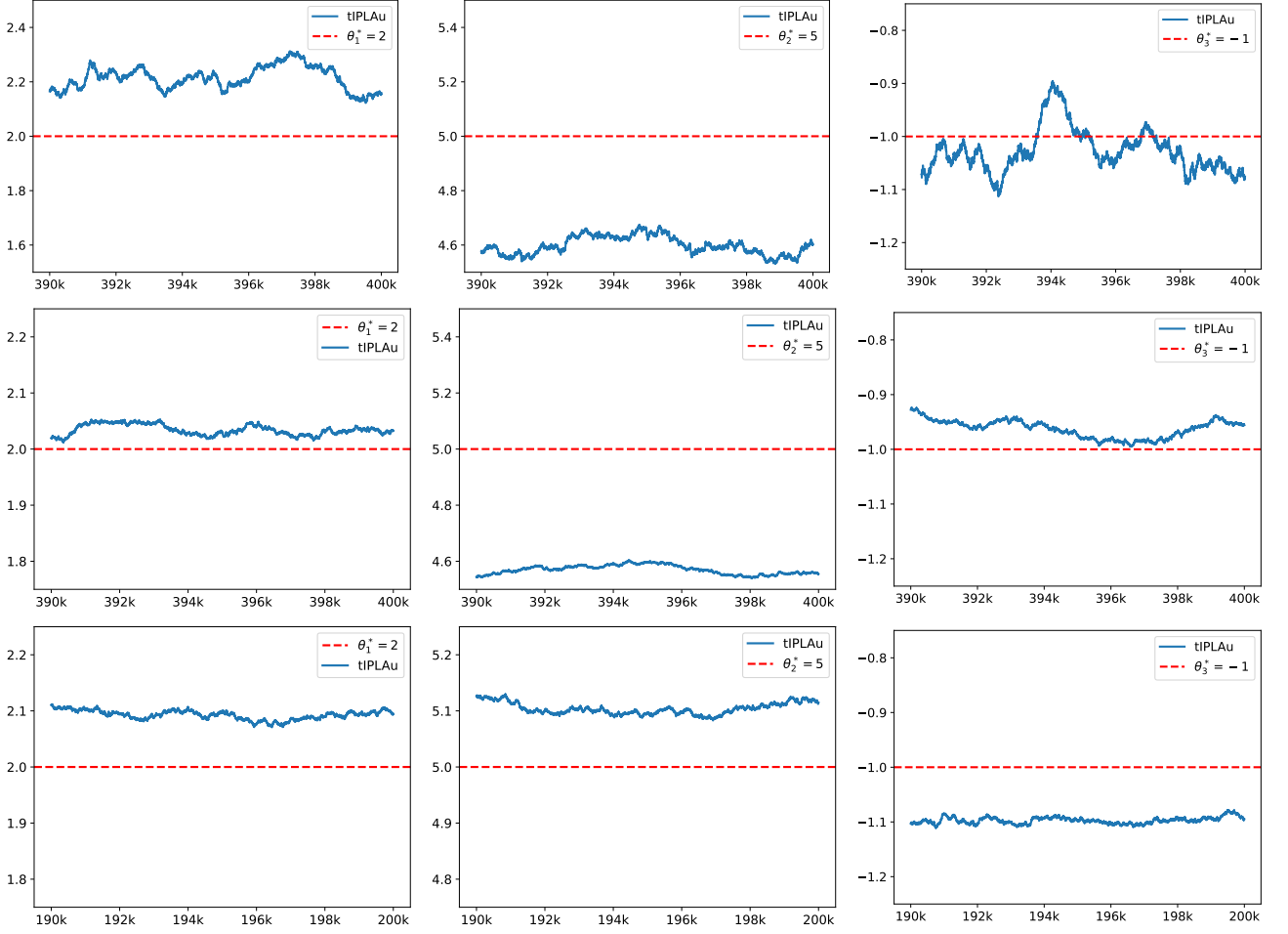
13

Figure 2: The last 10k iterates from tIPLAu's performance as shown in Fig 1. Each column corresponds to one of the component of the true parameter $\theta^* = [2, 5, -1]$. The top row corresponds to $N = 10$, the middle row to $N = 100$, and the bottom row to $N = 1000$

Now the potential $U(\theta, x)$ can be written as

$$U(\theta, x) = |x|^{4m} + (|x|^{2m} + 1)(|\theta|^{2m} + 1) + |\theta|^{4m} + |x|^4 + (|x|^2 + 1)(|\theta|^2 + 1) + |\theta|^4, \tag{25}$$

with the following gradient structure

$$\nabla_x U = 4m|x|^{4m-2}x + 2m(|\theta|^{2m} + 1)|x|^{2m-2}x + 4|x|^2 x + 2(|\theta|^2 + 1)x,$$
$$\nabla_\theta U = 4m|\theta|^{4m-2}\theta + 2m(|x|^{2m} + 1)|\theta|^{2m-2}\theta + 4|\theta|^2\theta + 2(|x|^2 + 1)\theta,$$

In light of Remark 5 and Remark 6 we verify that both A1 and A2 hold, in particular we have

$$|\nabla U(v) - \nabla U(v')| \leq 3m2^{4m} \left(1 + |v|^{4m-2} + |v'|^{4m-2}\right)|v - v'|,$$
$$\langle v - v', \nabla U(v) - \nabla U(v') \rangle \geq 2|v - v'|^2,$$

where $v = (\theta, x)$. To check that A3-i holds, we easily observe that

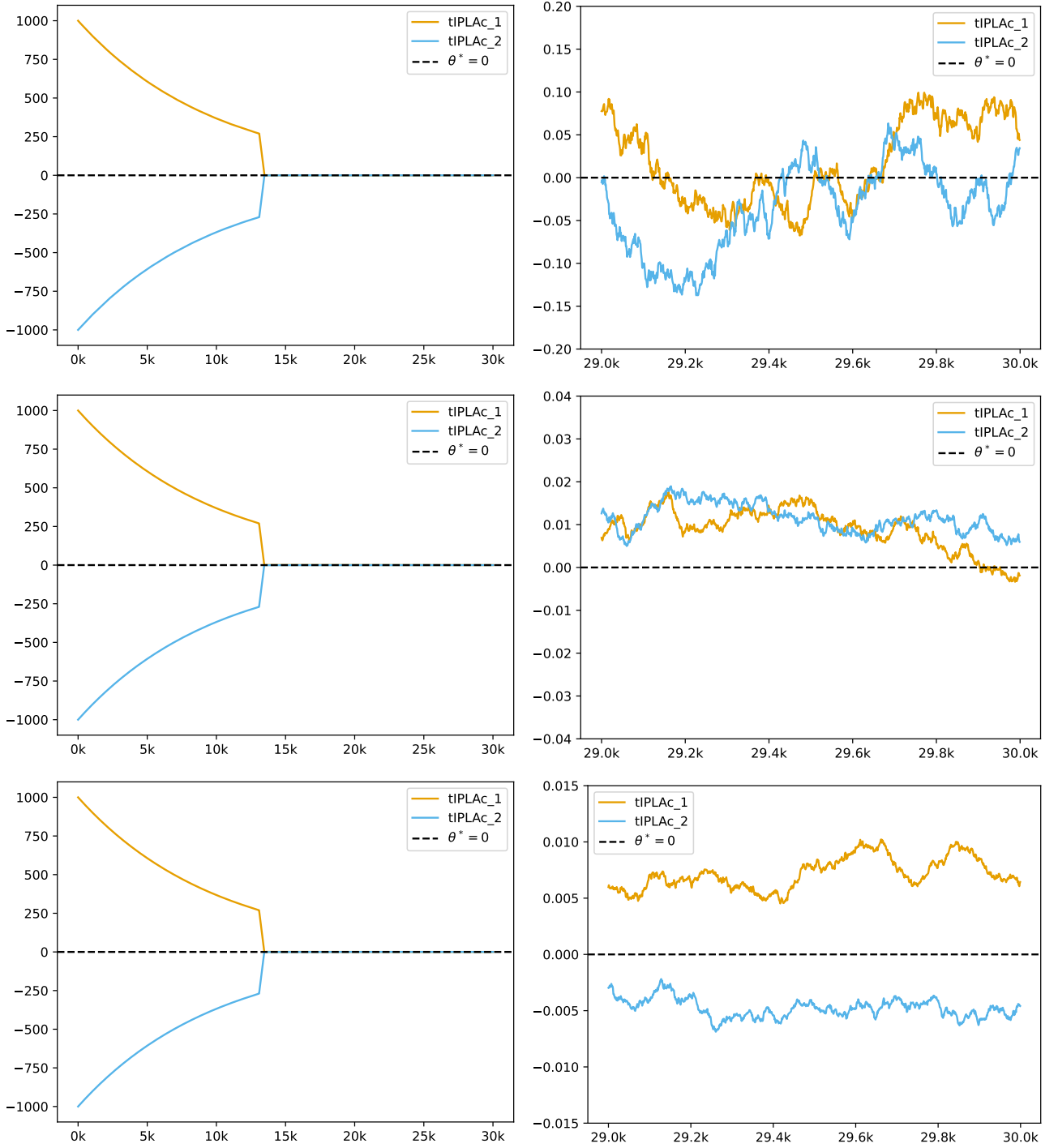$$\nabla_{\theta_i} U \theta_i \geq 2\theta_i^2 \text{ and } \nabla_{x_i} U x_i \geq 2x_i^2.$$

14

Figure 3: The performance of tIPLAc on the superlinear toy problem for $m = 15$. The top row corresponds to $N = 100$, the middle row to $N = 1000$, and the bottom row to $N = 10,000$. The left column displays the full iterates of the algorithm, while the right column focuses on the last 1000 iterations to examine any residual bias in the limiting behavior

15

Experiment details. We set $m = 15$ to a achieve a highly superlinear regime, where terms of the 60th order appear. For the experiment, we choose $d^x = 100$ and $d^\theta = 2$, with the true parameter θ^* known to be the zero vector $(0,0)$ to test parameter convergence. We run tIPLAc, for $M = 30$k iterations, using $N = 100, 1000, 10000$ particles and a stepsize of $\lambda = 0.0001$. The parameter θ_0 is initialized on purpose from a very high deterministic value $[1000, -1000]$, while the particles $X_0^{i,N}$ are drawn from a Gaussian distribution with randomized mean over the interval $(-100, 100)$ for each coordinate, and covariance matrix $R = 10I_{d^x}$.

As shown in Fig. 3, we observe that as N increases, the convergence time to the true value remains unchanged, contrary to the results from the tIPLAu algorithm. This is explained by the absence of N in the exponent of the second error term in Theorem 2. Furthermore, focusing on the second column, which displays the last 1000 iterations of each simulation, we notice a significant decrease of the residual bias of the algorithm once convergence has been achieved. This aligns with our theoretical finding, where we observe an one order of magnitude reduction in bias corresponding to a two order of magnitude increase in N. Additionally, it is evident that as N increases, the iterates of the algorithm generate less noisy Markov chains, confirming that N acts as the inverse temperature parameter in our setting (Fig. 4).
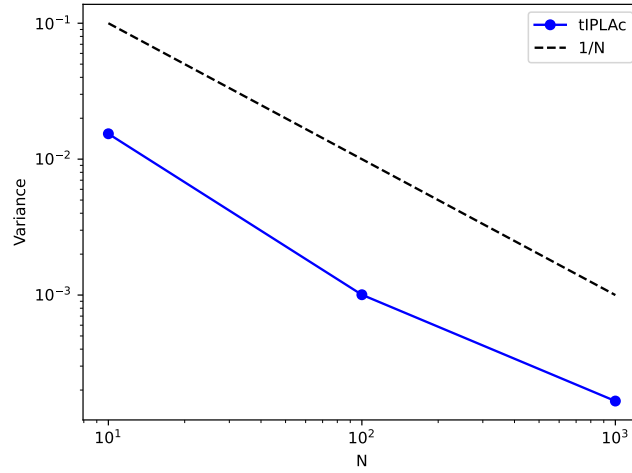

Figure 4: The convergence rate of the variance of the parameter estimates produced by tIPLAc over 100 Monte Carlo runs for each $N \in (10, 100, 100)$. We verify that the $\mathcal{O}\left(1/N\right)$ convergence rate holds for the second moments as suggested by our theoretical results

Furthermore, we deploy tIPLAc in par with the other aforementioned interacting particles based algorithms, that is IPLA [1], PGD [10] and SOUL [3] in a moderate superlinear setting with $m = 1$ and $N = 2$ number of particles. Once again we choose $d^x = 100$ and $d^\theta = 2$, with the true parameter θ^* known to be the zero vector $(0,0)$ to test parameter convergence. We run all all the algorithms for $M = 3000$ iterations and a stepsize of $\lambda = 0.0001$. The parameter θ_0 is initialized from a deterministic value $[3.28345, -3.28345]$, while the particles $X_0^{i,N}$ are drawn from a Gaussian distribution with randomized mean over the interval $(-0.1, 0.1)$ for each coordinate, and covariance matrix $R = 0.1I_{d^x}$.

As illustrated in Fig 5, tIPLAc demonstrates rapid convergence to the true parameter. In contrast, the non-tamed algorithms exhibit oscillatory behavior, ultimately diverging to infinity. This divergence arises due to the lack of theoretical guarantees in handling superlinearities, which are intrinsic to the tIPLA algorithms. These observations align with existing literature on unadjusted Langevin-based algorithms, where it is well-established that, in a superlinear setting, the iterates tend to diverge given sufficiently poor initial conditions. In this simple case, although the initial condition is only an order of magnitude away from the ground truth, divergence is still observed. Specifically, SOUL diverges as early as the third iteration, while IPLA and PGD exhibit divergence by the fifth iteration.

For the subsequent simulations, we set $m = 6$ to a achieve a mild superlinear regime with 12th order terms. We compare tIPLAu and tIPLAc under high dimensional settings. In particular we examine two cases per algorithm, namely $d^\theta = d^x = 100$ and $d^\theta = d^x = 1000$, with the true parameter θ^* known to be the zero vector to test parameter convergence. For tIPLAu, we run $M = 50$k iterations, using $N = 50$ particles and a stepsize of $\lambda = 0.0001$. Respectively tIPLAc is run, for $M = 20$k iterations, using $N = 50$ particles and a stepsize of $\lambda = 0.0001$. The parameter θ is a d^θ-dimensional vector where each coordinate is independently initialized to either 1000 or -1000, while the particles $X_0^{i,N}$ are drawn from a Gaussian distribution with randomized mean over the interval $(-10, 10)$ for each coordinate, and covariance matrix $R = 10I_{d^x}$.

16

As illustrated in Fig. 6, focusing on the first two rows, tIPLAc converges to the true parameter significantly faster than tIPLAu in both settings. This aligns with our theoretical results, as the convergence rate of tIPLAu deteriorates with increasing N due to time-scaling effects, whereas tIPLAc remains unaffected. Additionally, both algorithms exhibit slower convergence when the problem's dimensionality increases by an order of magnitude, consistent with the established Corollaries.

Examining the last two rows, which display the final 1500 iterations, we observe that the residual bias of tIPLAu is more sensitive to dimensionality, whereas tIPLAc remains robust. Moreover, tIPLAc consistently yields lower residual bias than tIPLAu. It is expected for coordinate-wise taming algorithms such as tIPLAc to outperform their uniform counterparts, as the latter indiscriminately modify all coordinates, failing to distinguish between stable and exploding ones. When the additional smoothness Assumption 3 holds, tIPLAc is preferable to tIPLAu.
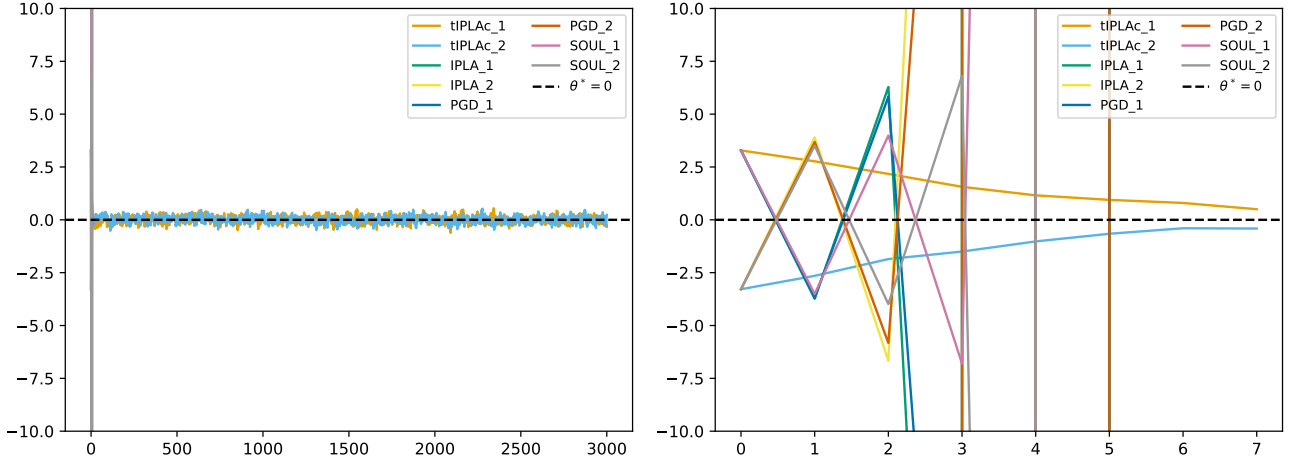

Figure 5: The performance of tIPLAc, IPLA, PGD and SOUL on the superlinear toy problem for $m = 1$ and $N = 100$. The left column displays the full iterates of the algorithms, while the right column focuses on the first 8 iterations where the explosive behavior of the non-tamed algorithms is observed

4.3. Mixed term - Toy example

To demonstrate the applicability of tIPLAc under assumption A3-ii, we consider the example from Section 4.2 for $m = 1$ but with an additional cross term $x\theta$,

$$k(\theta) = \int_{\mathbb{R}^{d^x}} \exp\left(-x\theta - |x|^4 - (|x|^2 + 1)(|\theta|^2 + 1) - |\theta|^4\right) dx.$$

Hence the potential $U(\theta, x) = x\theta + |x|^4 + (|x|^2 + 1)(|\theta|^2 + 1) + |\theta|^4$ has the following gradient structure

$$\nabla_x U = 4|x|^2 x + 2(|\theta|^2 + 1)x + \theta \text{ and } \nabla_\theta U = 4|\theta|^2\theta + 2(|x|^2 + 1)\theta + x.$$

One immediately checks that A1 holds, and in view of Remark 7, A2 is obtained as well, that is

$$|\nabla U(v) - \nabla U(v')| \le 52\left(1 + |v|^3 + |v'|^3\right)|v - v'|,$$
$$\langle v - v', \nabla U(v) - \nabla U(v')\rangle \ge |v - v'|^2,$$

where $v = (\theta, x)$. However the gradient of the potential $U(\theta, x)$ fails to satisfy A3-i, as shown below

$$(\nabla_x U)_i \, x_i = 4|x|^2 x_i^2 + 2(|\theta|^2 + 1)x_i^2 + x_i\theta_i \ge 2x_i^2 + x_i\theta_i,$$

17

since the product $x_i \theta_i$ cannot be controlled due to the fact that θ takes values in an unbounded domain. Nevertheless, we show that A3-ii holds with parameters $\mu = 3$, $\rho = 1/2$ and $b = 0$. Indeed

$$(\nabla_x U)_i \, x_i \geq 2x_i^2 - \frac{x_i^2}{2} - \frac{\theta_i^2}{2} = \frac{3}{2}x_i^2 - \frac{1}{2}\theta_i^2.$$

Experiment details. We choose $d^x = d^\theta = 10$, with the true parameter θ^* being unknown. We run tIPLAc, for $M = 20$k iterations, using $N = 100, 1000, 10000$ particles and a stepsize of $\lambda = 0.0001$. The parameter θ_0 is initialized from a deterministic value $[10,000, -10,000, 7500, -7500, 5000, -5000, 2500, -2500, 1000, -1000]$, while the particles $X_0^{i,N}$ are drawn from a Gaussian distribution with randomized mean over the interval $(-100, 100)$ for each coordinate, and covariance matrix $R = 10I_{d^x}$.

As shown in Fig. 7, we observe that as N increases, the convergence time to the true value remains unchanged, which agrees with our observation in Fig. 3. Furthermore, focusing on the second column, which displays the last 500 iterations of each simulation, we notice again that the residual bias of the algorithm, once convergence has been achieved, decreases significantly as the number of particles increases.
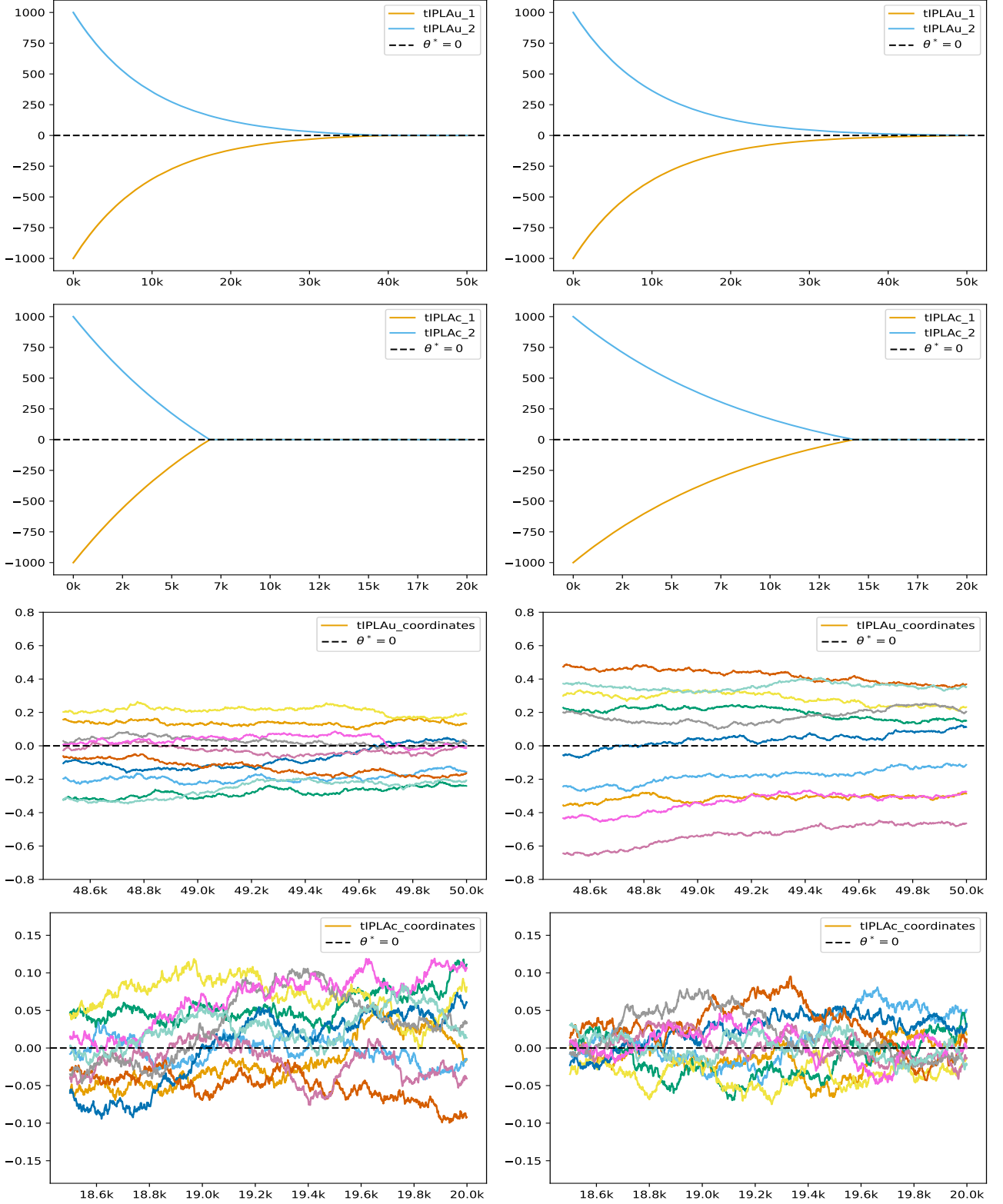
Figure 6: Performance comparison between tIPLAu and tIPLAc on the superlinear toy problem. The left column corresponds to $d^x = d^\theta = 100$, and the right column to $d^x = d^\theta = 1000$. The first two rows display the full iterates of the algorithm, while the last two rows focus on the last 1500 iterations to examine any residual bias in the limiting behavior
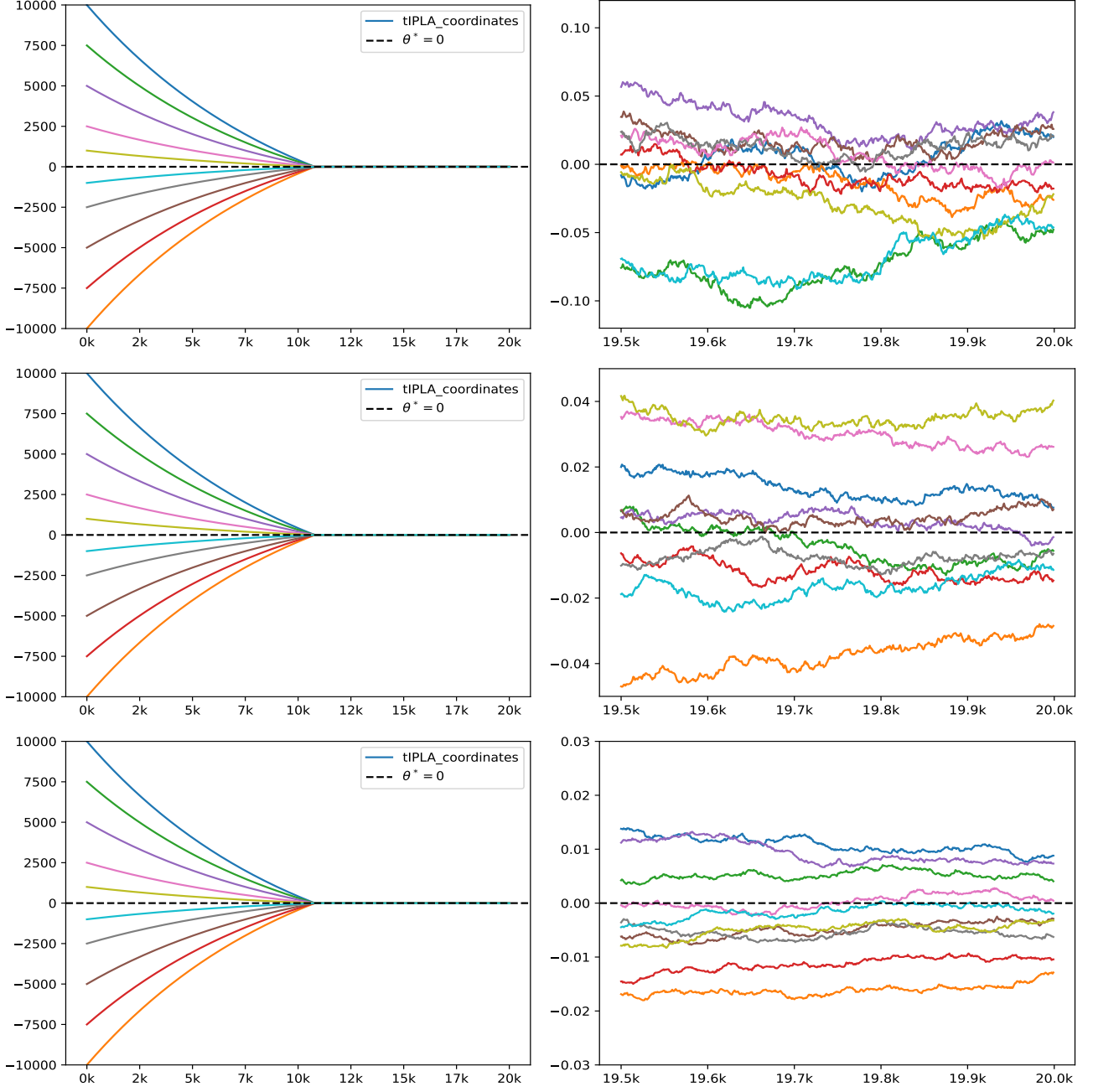
19

Figure 7: The performance of tIPLAc on the superlinear toy problem under assumption A3-ii. The top row corresponds to $N = 100$, the middle row to $N = 1000$, and the bottom row to $N = 10,000$. The left column displays the full iterates of the algorithm, while the right column focuses on the last 500 iterations to examine any residual bias in the limiting behavior

5. References

[1] Akyildiz, Ö.D., Crucinio, F.R., Girolami, M., Johnston, T., Sabanis, S.: Interacting Particle Langevin Algorithm for Maximum Marginal Likelihood Estimation, 2023, Preprint arXiv:2303.13429. https://doi.org/10.48550/arXiv.2303.13429

[2] Brosse, N., Durmus, A., Moulines, 'E., Sabanis, S.: The tamed unadjusted Langevin algorithm, Stoch. Proc. App. 129 (2019) 3638-3663. https://doi.org/10.1016/j.spa.2018.10.002

[3] De Bortoli, V., Durmus, A., Pereyra, M., Vidal, A.F.: Efficient stochastic optimisation by unadjusted Langevin Monte Carlo, Stat. Comput. 31 (2021). https://doi.org/10.1007/s11222-020-09986-y

[4] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm, J. R. Stat. Soc. Ser. B Methodol. 39 (1977) 1-38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

[5] Durmus, A., Moulines, 'E.: Nonasymptotic convergence analysis for the unadjusted Langevin algorithm, Ann. Appl. Probab. 27 (2017) 1551-1587. https://doi.org/10.1214/16-AAP1238

[6] Durmus, A., Moulines, 'E.: High-dimensional Bayesian inference via the unadjusted Langevin algorithm, Bernoulli 25 (2019) 2854-2882. https://doi.org/10.3150/18-BEJ1073

[7] Hutzenthaler, M., Martin, A., Kloeden, P.E.: Strong and weak divergence in finite time of Euler's method for stochastic differential equations with non-globally Lipschitz continuous coefficients, Proc. R. Soc. A Math. Phys. Eng. Sci. 467 (2010) 1563-1576. https://doi.org/10.1098/rspa.2010.0348

[8] Hutzenthaler, M., Jentzen, A., Kloeden, P.E.: Strong convergence of an explicit numerical method for SDEs with nonglobally Lipschitz continuous coefficients, Ann. Appl. Probab. 22 (2012). https://doi.org/10.1214/11-AAP803

[9] Kumar, C., Sabanis, S.: On explicit approximations for Lévy driven SDEs with super-linear diffusion coefficients, Electron. J. Probab. 22 (2017) 1-19. https://doi.org/10.1214/17-EJP89

[10] Kuntz, J., Lim, J.N., Johansen, A.M.: Particle algorithms for maximum likelihood training of latent variable models, Proc. 26th Int. Conf. Artif. Intell. Stat. 206 (2023) 5134-5180. https://proceedings.mlr.press/v206/kuntz23a.html

[11] Lim, D.Y., Sabanis, S.: Polygonal Unadjusted Langevin Algorithms: Creating stable and efficient adaptive algorithms for neural networks, J. Mach. Learn. Res. 25 (2024) 1-52. http://jmlr.org/papers/v25/22-0796.html

[12] Lim, D.Y., Neufeld, A., Sabanis, S., Zhang, Y.: Non-asymptotic estimates for TUSLA algorithm for non-convex learning with applications to neural networks with ReLU activation function, IMA J. Numer. Anal. 44 (2024) 1464-1559. https://doi.org/10.1093/imanum/drad038

[13] Lim, D.Y., Neufeld, A., Sabanis, S., Zhang, Y.: Langevin dynamics based algorithm e-THϵO POULA for stochastic optimization problems with discontinuous stochastic gradient, Math. Oper. Res. (2024). https://doi.org/10.1287/moor.2022.0307

[14] Lovas, A., Lytras, I., R'asonyi, M., Sabanis, S.: Taming Neural Networks with TUSLA: Nonconvex Learning via Adaptive Stochastic Gradient Langevin Algorithms, SIAM J. Math. Data Sci. 5 (2023) 323-345. https://doi.org/10.1137/22M1514283

[15] Mao, X.: Stochastic Differential Equations and Applications, 2nd edn. Elsevier, Hoboken, 2008.

[16] McCulloch, C.E.: Maximum Likelihood Algorithms for Generalized Linear Mixed Models, J. Am. Stat. Assoc. 92 (1997) 162-170. https://doi.org/10.1080/01621459.1997.10473613

[17] Prato, G.: An Introduction to Infinite-Dimensional Analysis, 1st edn. Springer, Berlin, 2006.

[18] Sabanis, S.: A note on tamed Euler approximations, Electr. Commun. Probab. 18 (2013) 1-10. https://doi.org/10.1214/ECP.v18-2824

[19] Sabanis, S.: Euler approximations with varying coefficients: The case of superlinearly growing diffusion coefficients, Ann. Appl. Probab. (2016). https://doi.org/10.1214/15-AAP1140

[20] Sabanis, S., Zhang, Y.: Higher order Langevin Monte Carlo algorithm, Electron. J. Stat. 13 (2019) 3805-3850. https://doi.org/10.1214/19-EJS1615

[21] Sabanis, S., Zhang, Y.: On explicit order 1.5 approximations with varying coefficients: The case of super-linear diffusion coefficients, J. Complex. 50 (2019) 84-115. https://doi.org/10.1016/j.jco.2018.09.004

[22] Saumard, A., Wellner, J.A.: Log-concavity and strong log-concavity: A review, Statist. Surv. 8 (2014) 45-114. https://doi.org/10.1214/14-SS107

[23] Wei, G.C.G., Tanner, M.A.: A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms, J. Am. Stat. Assoc. 85 (1990) 699-704. https://doi.org/10.2307/2290005

A. Proofs of Section 3

A.1. Uniform moments bounds

Lemma 1. *Consider either the dynamics given by* (1)-(2) *or* (3)-(4) *and let Assumptions A1-A2 hold, then there exists a constant $C > 0$ such that*

$$\sup_{t \geq 0} \mathbb{E}\left[|\mathcal{Z}_t^{\cdot,N}|^2\right] \leq C.$$

Proof. Let us first consider the rescaled dynamics as given by (12) for the original system of equations (1)-(2). The square norm of these dynamics is expressed as

$$|\mathcal{Z}_t^{c,N}|^2 = |\vartheta_t^{c,N}|^2 + \frac{1}{N}\sum_{i=1}^N |\mathcal{X}_t^{i,c,N}|^2.$$

Using standard arguments involving stopping times, Grönwall's lemma and Fatou's lemma, we obtain the existence of a constant c, which depends on time, such that $\sup_{0 \leq t \leq T} \mathbb{E}\left[|\mathcal{Z}_t^{c,N}|^2\right] \leq c$ for any $T > 0$. Furthermore, by applying Itô's formula, we derive

$$
\begin{aligned}
|\mathcal{Z}_t^{c,N}|^2 &\leq |\vartheta_0^{c,N}|^2 - 2\int_0^t \langle \vartheta_s^{\cdot,N}, \frac{1}{N}\sum_{i=1}^N h^\theta(\mathcal{V}_s^{i,c,N})\rangle ds + \frac{2d^\theta t}{N} + 2\sqrt{\frac{2}{N}}\int_0^t \vartheta_s^{c,N} dB_s^{0,N} + 2d^x t \\
&\quad + \frac{1}{N}\sum_{i=1}^N |\mathcal{X}_0^{i,c,N}|^2 - \frac{2}{N}\sum_{i=1}^N \int_0^t \langle \mathcal{X}_s^{i,c,N}, h^x(\mathcal{V}_s^{i,c,N})\rangle ds + \frac{2\sqrt{2}}{N}\sum_{i=1}^N \int_0^t \mathcal{X}_s^{i,c,N} dB_s^{i,N} ds \\
&\leq |\mathcal{Z}_0^{c,N}|^2 - 2\int_0^t \frac{1}{N}\sum_{i=1}^N \langle \mathcal{V}_s^{i,c,N}, h(\mathcal{V}_s^{i,c,N})\rangle ds + 2(d^\theta/N + d^x)t \\
&\quad + 2\sqrt{\frac{2}{N}}\int_0^t \vartheta_s^{c,N} dB_s^{0,N} + \frac{2\sqrt{2}}{N}\sum_{i=1}^N \int_0^t \mathcal{X}_s^{i,c,N} dB_s^{i,N} ds \\
&\leq |\mathcal{Z}_0^{c,N}|^2 - \mu \int_0^t |\mathcal{Z}_s^{c,N}|^2 ds + 2bt + 2(d^\theta/N + d^x)t \\
&\quad + 2\sqrt{\frac{2}{N}}\int_0^t \vartheta_s^{c,N} dB_s^{0,N} + \frac{2\sqrt{2}}{N}\sum_{i=1}^N \int_0^t \mathcal{X}_s^{i,c,N} dB_s^{i,N} ds.
\end{aligned}
$$

Taking the expectation on both sides, we obtain

$$\mathbb{E}\left[|\mathcal{Z}_t^{c,N}|^2\right] \leq \mathbb{E}\left[|\mathcal{Z}_0^{c,N}|^2\right] - \mu \int_0^t \mathbb{E}\left[|\mathcal{Z}_s^{c,N}|^2\right] ds + 2(b + d^\theta/N + d^x)t.$$

$$\frac{d}{dt}\exp(\mu t)\mathbb{E}\left[|\mathcal{Z}_t^{c,N}|^2\right] \leq \exp(\mu t)C,$$

which, upon integration, leads to

$$\mathbb{E}\left[|\mathcal{Z}_t^{c,N}|^2\right] \leq C/\mu \Rightarrow \sup_{t \geq 0} \mathbb{E}\left[|\mathcal{Z}_t^{c,N}|^2\right] \leq C,$$

as C is a constant independent of time. The corresponding result regarding equations (3)-(4) follows by going through the same steps, with the calculations differing only up to a constant in the SDEs coefficients. □

A.2. Uniformly tamed scheme

A.2.1 Key quantities for the proof of the main Lemmas.

The following definitions refer to the rescaled dynamics (15)-(16) of the algorithm tIPLAu and its continuous time interpolations (17)-(18)

$$Z_n^{\lambda,u} = \left(\theta_{n+1}^{\lambda,u}, N^{-1/2} X_{n+1}^{1,\lambda,u}, \ldots, N^{-1/2} X_{n+1}^{N,\lambda,u}\right), \tag{26}$$

$$\overline{Z}_t^{\lambda,u} = \left(\overline{\theta}_t^{\lambda,u}, N^{-1/2}\overline{X}_t^{1,\lambda,u}, \ldots, N^{-1/2}\overline{X}_t^{N,\lambda,u}\right), \tag{27}$$

$$\mathcal{Z}_{\lambda t}^{u,N} = \left(\vartheta_{\lambda t}^{u,N}, N^{-1/2}\mathcal{X}_{\lambda t}^{1,u,N}, \ldots, N^{-1/2}\mathcal{X}_{\lambda t}^{N,u,N}\right). \tag{28}$$

A.2.2 Moment and increment bounds

Lemma 2. *Let A1, A2 and A4 hold. Then, for any $0 \leq \lambda < N^p/(4\mu)$, it holds that*

$$\mathbb{E}\left[|Z_n^{\lambda,u}|^2\right] \leq C_{|z_0|,\mu,b}(1 + d^\theta/N + d^x),$$

for a constant $C > 0$, independent of N, n, λ, d^x and d^θ, given explicitly within the proof.

Proof. Consider the rescaled iterates as described in equation (26), which expands as

$$\left|Z_{n+1}^{\lambda,u}\right|^2 = \left|\theta_{n+1}^{\lambda,u}\right|^2 + \frac{1}{N}\sum_{i=1}^N \left|X_{n+1}^{i,\lambda,u}\right|^2$$

$$= \left|\theta_n^{\lambda,u} - \frac{\lambda}{N^{p+1}}\sum_{i=1}^N h_{\lambda,u}^\theta(\theta_n^{\lambda,u}, X_n^{i,\lambda,u})\right|^2 + \frac{2\lambda}{N^{p+1}}|\xi_{n+1}^{(0)}|^2$$

$$+ 2\sqrt{\frac{2\lambda}{N^{p+1}}}\left\langle \theta_n^{\lambda,u} - \frac{\lambda}{N^{p+1}}\sum_{i=1}^N h_{\lambda,u}^\theta(\theta_n^{\lambda,u}, X_n^{i,\lambda,u}), \xi_{n+1}^{(0)}\right\rangle$$

$$+ \frac{1}{N}\sum_{i=1}^N \left(\left|X_n^{i,\lambda,u} - \frac{\lambda}{N^p}h_{\lambda,u}^x(\theta_n^{\lambda,u}, X_n^{i,\lambda,u})\right|^2 + \frac{2\lambda}{N^p}|\xi_{n+1}^{(i)}|^2\right.$$

$$\left. + 2\sqrt{\frac{2\lambda}{N^p}}\left\langle X_n^{i,\lambda,u} - \frac{\lambda}{N^p}h_{\lambda,u}^x(\theta_n^{\lambda,u}, X_n^{i,\lambda,u}), \xi_{n+1}^{(i)}\right\rangle\right).$$

Taking the conditional expectation of both sides with respect to the filtration generated by $Z_n^{\lambda,u}$ the cross terms vanish to 0

due to the independence between the $\xi_{n+1}^{(i)}$'s and $Z_n^{\lambda,u}$. This yields

$$
\begin{aligned}
\mathbb{E}\left[\left|Z_{n+1}^{\lambda,u}\right|^2 \mid Z_n^{\lambda,u}\right] = {}& \mathbb{E}\left[\left|\theta_n^{\lambda,u}\right|^2 \mid Z_n^{\lambda,u}\right] - \frac{2\lambda}{N^{p+1}}\sum_{i=1}^{N}\mathbb{E}\left[\langle \theta_n^{\lambda,u}, h_{\lambda,u}^\theta(\theta_n^{\lambda,u}, X_n^{i,\lambda,u})\rangle \mid Z_n^{\lambda,u}\right] \\
& + \frac{\lambda^2}{N^{2(p+1)}}\mathbb{E}\left[\left|\sum_{i=1}^{N} h_{\lambda,u}^\theta(\theta_n^{\lambda,u}, X_n^{i,\lambda,u})\right|^2 \mid Z_n^{\lambda,u}\right] + \frac{2\lambda d^\theta}{N^{p+1}} \\
& + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left|X_n^{i,\lambda,u}\right|^2 \mid Z_n^{\lambda,u}\right] \\
& - \frac{2\lambda}{N^{p+1}}\sum_{i=1}^{N}\mathbb{E}\left[\langle X_n^{i,\lambda}, h_{\lambda,u}^x(\theta_n^{\lambda,u}, X_n^{i,\lambda,u})\rangle \mid Z_n^{\lambda,u}\right] \\
& + \frac{\lambda^2}{N^{2p+1}}\sum_{i=1}^{N}\mathbb{E}\left[|h_{\lambda,u}^x(\theta_n^{\lambda,u}, X_n^{i,\lambda,u})|^2 \mid Z_n^{\lambda,u}\right] + \frac{2\lambda d^x}{N^p}.
\end{aligned}
$$

Next, using the elementary inequality $(t_1 + \ldots + t_m)^p \le m^{p-1}(t_1^p + \ldots, t_m^p)$ along with the fact that all of the terms inside the conditional expectations are measurable, we simplify the expression to

$$
\begin{aligned}
\mathbb{E}\left[\left|Z_{n+1}^{\lambda,u}\right|^2 \mid Z_n^{\lambda,u}\right] \le {}& |\theta_n^{\lambda,u}| + \frac{1}{N}\sum_{i=1}^{N}|X_n^{i,\lambda,u}|^2 + \frac{2\lambda d^\theta}{N^{p+1}} + + \frac{2\lambda d^x}{N^p} \\
& - \frac{2\lambda}{N^{p+1}}\sum_{i=1}^{N}\langle \theta_n^\lambda, h_{\lambda,u}^\theta(\theta_n^{\lambda,u}, X_n^{i,\lambda,u})\rangle \\
& - \frac{2\lambda}{N^{p+1}}\sum_{i=1}^{N}\langle X_n^{i,\lambda,u}, h_{\lambda,u}^x(\theta_n^{\lambda,u}, X_n^{i,\lambda,u})\rangle \\
& + \frac{\lambda^2}{N^{2p+1}}\sum_{i=1}^{N}|h_\lambda^\theta(\theta_n^{\lambda,u}, X_n^{i,\lambda,u})|^2 + \frac{\lambda^2}{N^{2p+1}}\sum_{i=1}^{N}|h_{\lambda,u}^x(\theta_n^{\lambda,u}, X_n^{i,\lambda,u})|^2.
\end{aligned}
$$

Recalling that the two drift terms can be written together as $(h_{\lambda,u}^\theta, h_{\lambda,u}^x) = h_{\lambda,u}$, leads to

$$
\begin{aligned}
\mathbb{E}\left[\left|Z_{n+1}^{\lambda,u}\right|^2 \mid Z_n^{\lambda,u}\right] \le {}& \left|Z_n^{\lambda,u}\right|^2 - \frac{2\lambda}{N^{p+1}}\sum_{i=1}^{N}\langle V_n^{i,\lambda,u}, h_{\lambda,u}(V_n^{i,\lambda,u})\rangle \\
& + \frac{\lambda^2}{N^{2p+1}}\sum_{i=1}^{N}|h_{\lambda,u}(V_n^{i,\lambda,u})|^2 + \frac{2\lambda}{N^p}(d^\theta/N + d^x).
\end{aligned}
$$

Now, by applying Properties 1 and 3 of the taming function, we simplify further

$$
\begin{aligned}
\mathbb{E}\left[\left|Z_{n+1}^{\lambda,u}\right|^2 \mid Z_n^{\lambda,u}\right] \le {}& \left|Z_n^{\lambda,u}\right|^2 - \frac{\mu\lambda}{N^{p+1}}\sum_{i=1}^{N}\left|V_n^{i,\lambda,u}\right|^2 + \frac{2\lambda b}{N^p} \\
& + \frac{2\lambda^2\mu^2}{N^{2p+1}}\sum_{i=1}^{N}\left|V_n^{i,\lambda,u}\right|^2 + \frac{2\lambda^2\lambda^{-1}N^p}{N^{2p}} + \frac{2\lambda}{N^p}(d^\theta/N + d^x) \\
\le {}& \left(1 - \frac{\lambda\mu}{N^p} + \frac{2\lambda^2\mu^2}{N^{2p}}\right)\left|Z_n^{\lambda,u}\right|^2 + \frac{2\lambda}{N^p}\left(b + 1 + d^\theta/N + d^x\right).
\end{aligned}
$$

Iterating this inequality, and assuming the restriction $\lambda < N^p/(4\mu)$, yields the desired bound

$$
\begin{aligned}
\mathbb{E}\left[\left|Z_{n+1}^{\lambda,u}\right|^2\right] &\leq \left(1 - \frac{\mu\lambda}{2N^p}\right)^n \mathbb{E}\left[\left|Z_0^{\lambda,u}\right|^2\right] + \frac{1-(1-\mu\lambda/2)^n}{\mu\lambda/2N^p} \frac{2\lambda}{N^p}(b+1+d^\theta/N+d^x) \\
&\leq \mathbb{E}\left[\left|Z_0^{\lambda,u}\right|^2\right] + \frac{4}{\mu}(b+1+d^\theta/N+d^x) \\
&\leq \left(\mathbb{E}\left[\left|Z_0^{\lambda,u}\right|^2\right] + \frac{4}{\mu}(b+1)\right)(1+d^\theta/N+d^x) \\
&\leq C_{|z_0|,\mu,b}(1+d^\theta/N+d^x).
\end{aligned}
$$

$\square$

Lemma 3. *Let A1, A2 and A4 hold. Then, for any $0 \leq \lambda < N^p/(4\mu)$ and $q \in [2, 2\ell+1] \cap \mathbb{N}$, it holds that,*

$$
\mathbb{E}\left[|Z_n^{\lambda,u}|^{2q}\right] \leq C_{|z_0|,b,q,\mu}(1+d^\theta/N+d^x)^q,
$$

for a constant $C > 0$, independent of N, n, λ, d^x and d^θ, given explicitly within the proof.

Proof. To make the following calculations clearer, we define the auxiliary processes

$$
\Delta_n^{\lambda,\theta} = \theta_n^{\lambda,u} - \frac{\lambda}{N^{p+1}} \sum_{i=1}^N h_{\lambda,u}^\theta\left(\theta_n^\lambda, X_n^{i,\lambda}\right), \ G_n^{\lambda,\theta} = \sqrt{\frac{2\lambda}{N^{p+1}}} \xi_{n+1}^{(0)},
$$

$$
\Delta_n^{\lambda,x,i} = X_n^{i,\lambda,u} - \frac{\lambda}{N^p} h_{\lambda,u}^x\left(\theta_n^\lambda, X_n^{i,\lambda}\right), \ G_n^{\lambda,x,i} = \sqrt{\frac{2\lambda}{N^p}} \xi_{n+1}^{(i)}.
$$

For the sake of simplicity in the calculations, we denote λ/N^p as λ from this point onwards. Recall from Lemma 2 that we have the bound

$$
A_n^\lambda := |\Delta_n^{\lambda,\theta}|^2 + \frac{1}{N}\sum_{i=1}^N |\Delta_n^{\lambda,i,x}|^2 \leq \left(1 - \frac{\lambda\mu}{2}\right)|Z_n^{\lambda,u}|^2 + 2\lambda C.
$$

Now, we define the auxiliary quantity

$$
B_n^\lambda = 2\langle \Delta_n^\theta, G_n^{\lambda,\theta}\rangle + \frac{2}{N}\sum_{i=1}^N \langle \Delta_n^{\lambda,x,i}, G_n^{\lambda,x,i}\rangle + |G_n^{\lambda,\theta}|^2 + \frac{1}{N}\sum_{i=1}^N |G_n^{\lambda,x,i}|^2.
$$

For the $2q$-th moment, we can write

$$
\begin{aligned}
|Z_{n+1}^{\lambda,u}|^{2q} &= \left(A_n^\lambda + B_n^\lambda\right)^q \\
&\leq (A_n^\lambda)^q + 2q(A_n^\lambda)^{q-1}B_n^\lambda + \sum_{k=2}^q \binom{q}{k}|A_n^\lambda|^{q-k}|B_n^\lambda|^k.
\end{aligned}
\tag{29}
$$

We will now handle each term separately

$$
\begin{aligned}
\mathbb{E}\left[(A_n^\lambda)^q|Z_n^{\lambda,u}\right] = (A_n^\lambda)^q &\leq \left(\left(1-\frac{\lambda\mu}{2}\right)|Z_n^{\lambda,u}|^2 + 2\lambda C\right)^q \\
&\leq \left(1+\frac{\lambda\mu}{4}\right)^{q-1}\left(1-\frac{\lambda\mu}{2}\right)^q |Z_n^{\lambda,u}|^{2q} + \left(1+\frac{4}{\lambda\mu}\right)^{q-1} 2^q\lambda^q C^q \\
&\leq \left(1-\frac{\lambda\mu}{4}\right)^{q-1}\left(1-\frac{\lambda\mu}{2}\right)|Z_n^{\lambda,u}|^{2q} + \left(\lambda+\frac{4}{\mu}\right)^{q-1}\lambda(2C)^q \\
&\leq r_q^\lambda |Z_n^{\lambda,u}|^{2q} + w_q^\lambda,
\end{aligned}
\tag{30}
$$

25

where $r_q^\lambda = (1 - \lambda\mu/4)^{q-1}(1 - \lambda\mu/2)$ and $w_q^\lambda = (\lambda + 4/\mu)^{q-1}\lambda(2C)^q$. Notice that we used the elementary equation $(r + s)^q \leq (1 + \epsilon)^{q-1}r^q + (1 + 1/\epsilon)^{q-1}s^q$ with $\epsilon = \lambda\mu/4$. Additionally for simplicity, we denote $d = d^\theta/N + d^x$. Now, similarly

$$
\begin{aligned}
\mathbb{E}\left[2q(A_n^\lambda)^{q-1}B_n^\lambda|Z_n^{\lambda,u}\right] = 2q(A_n^\lambda)^{q-1}\mathbb{E}\left[B_n^\lambda|Z_n^{\lambda,u}\right] &= 4q\lambda\left(d^\theta/N + d^x\right)(A_n^\lambda)^{q-1} \\
&\leq 4q\lambda\left(d^\theta/N + d^x\right)\left(r_{q-1}^\lambda|Z_n^{\lambda,u}|^{2(q-1)} + w_{q-1}^\lambda\right) \\
&\leq 4q\lambda d\left(r_{q-1}^\lambda|Z_n^{\lambda,u}|^{2(q-1)} + w_{q-1}^\lambda\right).
\end{aligned}
\tag{31}
$$

The third term in equation (29) can be expanded to

$$
\begin{aligned}
\sum_{k=2}^{q}\binom{q}{k}|A_n^\lambda|^{q-k}|B_n^\lambda|^k &= \sum_{m=0}^{q-2}\binom{q}{m+2}|A_n^\lambda|^{q-2-m}|B_n^\lambda|^{m+2} \\
&= \frac{q}{m+2}\frac{q-1}{m+1}\sum_{m=0}^{q-2}\binom{q-2}{m}|A_n^\lambda|^{q-2-m}|B_n^\lambda|^m|B_n^\lambda|^2 \\
&\leq q(q-1)\left(|A_n^\lambda| + |B_n^\lambda|\right)^{q-2}|B_n^\lambda|^2 \\
&\leq q(q-1)2^{q-3}|A_n^\lambda|^{q-2}|B_n^\lambda|^2 + q(q-1)2^{q-3}|B_n^\lambda|^q \\
&:= D + F.
\end{aligned}
\tag{32}
$$

Taking the expectation in equation (32) yields

$$
\mathbb{E}[D|Z_n^{\lambda,u}] = q(q-1)2^{q-3}|A_n^\lambda|^{q-2}\mathbb{E}\left[|B_n^\lambda|^2|Z_n^{\lambda,u}\right],
$$

where

$$
\begin{aligned}
&\mathbb{E}\left[|B_n^\lambda|^2|Z_n^{\lambda,u}\right] \\
&= \mathbb{E}\left[\left|2\langle\Delta_n^\theta, G_n^{\lambda,\theta}\rangle + \frac{2}{N}\sum_{i=1}^{N}\langle\Delta_n^{\lambda,x,i}, G_n^{\lambda,x,i}\rangle + |G_n^{\lambda,\theta}|^2 + \frac{1}{N}\sum_{i=1}^{N}|G_n^{\lambda,x,i}|^2\right|^2\Big|Z_n^{\lambda,u}\right] \\
&\leq 4\left(\mathbb{E}\left[4|\Delta_n^{\lambda,\theta}|^2|G_n^{\lambda,\theta}|^2|Z_n^{\lambda,u}\right] + \mathbb{E}\left[\frac{4}{N}\sum_{i=1}^{N}|\Delta_n^{\lambda,x,i}|^2|G_n^{\lambda,x,i}|^2|Z_n^{\lambda,u}\right]\right. \\
&\quad + \mathbb{E}\left[|G_n^{\lambda,x,i}|^4|Z_n^{\lambda,u}\right] + \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}|G_n^{\lambda,x,i}|^4|Z_n^{\lambda,u}\right]\Bigg) \\
&\leq 4\left(4|\Delta_n^{\lambda,\theta}|^2\mathbb{E}\left[|G_n^{\lambda,\theta}|^2|Z_n^{\lambda,u}\right] + \frac{4}{N}\sum_{i=1}^{N}|\Delta_n^{\lambda,x,i}|^2\mathbb{E}\left[|G_n^{\lambda,x,i}|^2|Z_n^{\lambda,u}\right]\right. \\
&\quad + \mathbb{E}\left[|G_n^{\lambda,x,i}|^4|Z_n^{\lambda,u}\right] + \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}|G_n^{\lambda,x,i}|^4|Z_n^{\lambda,u}\right]\Bigg).
\end{aligned}
$$

Recalling that each $G_n^{\lambda,\cdot,i}$ follows a Gaussian distribution, we subsequently derive

$$
\begin{aligned}
&\mathbb{E}\left[|B_n^\lambda|^2|Z_n^{\lambda,u}\right] \\
&\leq 4\left(4(2\lambda d^\theta/N)|\Delta_n^{\lambda,\theta}|^2 + \frac{4}{N}\sum_{i=1}^{N}(2\lambda d^x)|\Delta_n^{\lambda,x,i}|^2 + 3(2\lambda d^\theta/N)^2 + 3(2\lambda d^x)^2\right) \\
&\leq 16(2\lambda d)\left(|\Delta_n^{\lambda,\theta}|^2 + \frac{1}{N}\sum_{i=1}^{N}|\Delta_n^{\lambda,x,i}|^2\right) + 24(2\lambda d)^2 \\
&\leq 16(2\lambda d)A_n^\lambda + 24(2\lambda d)^2.
\end{aligned}
$$

Substituting this into the term for D, we get

$$
\begin{aligned}
\mathbb{E}[D|Z_n^{\lambda,u}] &= q(q-1)2^{q-3}|A_n^\lambda|^{q-2}\left(16(2\lambda d)A_n^\lambda + 24(2\lambda d)^2\right) \\
&= 16(2\lambda d)q(q-1)2^{q-3}\left(r_{q-1}^\lambda|Z_n^{\lambda,u}|^{2(q-1)} + w_{q-1}^\lambda\right) \\
&\quad + 24(2\lambda d)^2 q(q-1)2^{q-3}\left(r_{q-2}^\lambda|Z_n^{\lambda,u}|^{2(q-2)} + w_{q-2}^\lambda\right).
\end{aligned}
\tag{33}
$$

Additionally

$$
\begin{aligned}
\mathbb{E}[F|Z_n^{\lambda,u}] &= q(q-1)2^{q-3}|B_n^\lambda|^q \\
&= q(q-1)2^{q-3}\mathbb{E}\left[\left|2\langle\Delta_n^\theta, G_n^{\lambda,\theta}\rangle + \frac{2}{N}\sum_{i=1}^N\langle\Delta_n^{\lambda,x,i}, G_n^{\lambda,x,i}\rangle\right.\right. \\
&\quad \left.\left. +|G_n^{\lambda,\theta}|^2 + \frac{1}{N}\sum_{i=1}^N|G_n^{\lambda,x,i}|^2\right|^q|Z_n^{\lambda,u}\right] \\
&\leq q(q-1)2^{q-3}4^{q-1}\left(\mathbb{E}\left[2^q|\Delta_n^{\lambda,\theta}|^q|G_n^{\lambda,\theta}|^q|Z_n^{\lambda,u}\right]\right. \\
&\quad + \mathbb{E}\left[\frac{2^q}{N^q}\left|\sum_{i=1}^N|\Delta_n^{\lambda,x,i}||G_n^{\lambda,x,i}|\right|^q|Z_n^{\lambda,u}\right] + \mathbb{E}\left[|G_n^{\lambda,\theta}|^{2q}|Z_n^{\lambda,u}\right] \\
&\quad \left.+\mathbb{E}\left[\frac{1}{N^q}N^{q-1}\sum_{i=1}^N|G_n^{\lambda,x,i}|^{2q}|Z_n^{\lambda,u}\right]\right).
\end{aligned}
\tag{34}
$$

We handle the second term in equation (34) by applying the multinomial expansion

$$
\left|\sum_{i=1}^N|\Delta_n^{\lambda,x,i}||G_n^{\lambda,x,i}|\right|^q = \sum_{k_1+\ldots+k_N=q}\binom{q}{k_1,\ldots,k_N}\prod_{i=1}^N\left(|\Delta_n^{\lambda,x,i}||G_n^{\lambda,x,i}|\right)^{k_i},
$$

and using the fact that $\Delta_n^{\lambda,x,i}$ are $Z_n^{\lambda,u}-$measurable, we obtain

$$
\begin{aligned}
&\frac{2^q}{N^q}\mathbb{E}\left[\left|\sum_{i=1}^N|\Delta_n^{\lambda,x,i}|G_n^{\lambda,x,i}|\right|^q|Z_n^{\lambda,u}\right] \\
&= \frac{2^q}{N^q}\mathbb{E}\left[\sum_{k_1+\ldots+k_N=q}\binom{q}{k_1,\ldots,k_N}\prod_{i=1}^N\left(|\Delta_n^{\lambda,x,i}||G_n^{\lambda,x,i}|\right)^{k_i}|Z_n^{\lambda,u}\right] \\
&= \frac{2^q}{N^q}\sum_{k_1+\ldots+k_N=q}\binom{q}{k_1,\ldots,k_N}\prod_{i=1}^N|\Delta_n^{\lambda,x,i}|^{k_i}\prod_{i=1}^N\mathbb{E}\left[|G_n^{\lambda,x,i}|^{k_i}|Z_n^{\lambda,u}\right] \\
&= \frac{2^q}{N^q}\sum_{k_1+\ldots+k_N=q}\binom{q}{k_1,\ldots,k_N}\prod_{i=1}^N|\Delta_n^{\lambda,x,i}|^{k_i}(2\lambda d^x)^{\sum_{i=1}^N k_i/2}k_i!! \\
&\leq \frac{2^q q!!(2\lambda d^x)^{q/2}}{N^q}\sum_{k_1+\ldots+k_N=q}\binom{q}{k_1,\ldots,k_N}\prod_{i=1}^N|\Delta_n^{\lambda,x,i}|^{k_i} \\
&= \frac{2^q q!!(2\lambda d^x)^{q/2}}{N^q}\left(\sum_{i=1}^N|\Delta_n^{\lambda,x,i}|\right)^q \leq \frac{2^q q!!(2\lambda d^x)^{q/2}}{N^{q-1}}\left(\sum_{i=1}^N|\Delta_n^{\lambda,x,i}|^2\right)^{q/2} \\
&\leq 2^q q!!(2\lambda d^x)^{q/2}\left(\frac{1}{N}\sum_{i=1}^N|\Delta_n^{\lambda,x,i}|^2\right)^{q/2}.
\end{aligned}
$$

Substituting these results into inequality (34), we derive

$$\mathbb{E}[F|Z_n^{\lambda,u}] \leq q(q-1)2^{3q-5}\left(2^q q!!(2\lambda d^\theta/N)^{q/2}\left(|\Delta_n^{\lambda,\theta}|^2\right)^{q/2}\right.$$

$$\left. +2^q q!!(2\lambda d^x)^{q/2}\left(\frac{1}{N}\sum_{i=1}^N|\Delta_n^{\lambda,x,i}|^2\right)^{q/2} + (2q)!!(2\lambda d^\theta/N)^q + (2q)!!(2\lambda d^x)^q\right)$$

$$\leq q(q-1)2^{3q-5}\left(2^q q!!(2\lambda d)^{q/2}\left(|\Delta_n^{\lambda,\theta}|^2 + \frac{1}{N}\sum_{i=1}^N|\Delta_n^{\lambda,x,i}|^2\right)^{q/2} + 2(2q)!!(2\lambda d)^q\right)$$

$$\leq q(q-1)2^{4q-5}(2q)!!\left((2\lambda d)^{q/2}\left|A_n^\lambda\right|^{q/2} + (2\lambda d)^q\right)$$

$$\leq q(q-1)2^{4q-5}(2q)!!(2\lambda d)^{q/2}\left(r_{q/2}^\lambda|Z_n^{\lambda,u}|^q + w_{q/2}^\lambda\right) + q(q-1)2^{4q-5}(2q)!!(2\lambda d)^q. \tag{35}$$

Combining the bounds in (33)-(35) yields the following control over the third term in (29)

$$\mathbb{E}\left[\sum_{k=2}^q\binom{q}{k}|A_n^\lambda|^{q-k}|B_n^\lambda|^k\right] \leq 16(2\lambda d)q(q-1)2^{q-3}\left(r_{q-1}^\lambda|Z_n^{\lambda,u}|^{2(q-1)} + w_{q-1}^\lambda\right)$$

$$+ 24(2\lambda d)^2 q(q-1)2^{q-3}\left(r_{q-2}^\lambda|Z_n^{\lambda,u}|^{2(q-2)} + w_{q-2}^\lambda\right)$$

$$+ (2q)!!(2\lambda d)^{q/2}q(q-1)2^{4q-5}\left(r_{q/2}^\lambda|Z_n^{\lambda,u}|^q + w_{q/2}^\lambda\right)$$

$$+ (2q)!!(2\lambda d)^q q(q-1)2^{4q-5}. \tag{36}$$

Hence, we obtain the following bound for (29) via (30),(31) and (36)

$$\mathbb{E}\left[|Z_{n+1}|^{2q}|Z_n\right] \leq r_q^\lambda|Z_n^{\lambda,u}|^{2q} + w_q^\lambda$$

$$+ 4q\,(2\lambda d)\,(1 + 4(q-1)2^{q-3})\left(r_{q-1}^\lambda|Z_n^{\lambda,u}|^{2(q-1)} + w_{q-1}^\lambda\right)$$

$$+ 24(2\lambda d)^2 q(q-1)2^{q-3}\left(r_{q-2}^\lambda|Z_n^{\lambda,u}|^{2(q-2)} + w_{q-2}^\lambda\right)$$

$$+ (2q)!!(2\lambda d)^{q/2}q(q-1)2^{4q-5}\left(r_{q/2}^\lambda|Z_n^{\lambda,u}|^q + w_{q/2}^\lambda\right)$$

$$+ (2q)!!(2\lambda d)^q q(q-1)2^{4q-5}. \tag{37}$$

We consider $|Z_n^{\lambda,u}| \geq \sqrt{8d/\mu}\left\{(2q)!!q(q-1)2^{4q-5}\right\}^{1/2} \geq \sqrt{8d/\mu}\left\{(2q)!!q(q-1)2^{4q-5}\right\}^{1/q}$. Thus we compute

$$\mathbb{E}\left[|Z_{n+1}^{\lambda,u}|^{2q}|Z_n\right] \leq r_q^\lambda|Z_n^{\lambda,u}|^{2q} + w_q^\lambda$$

$$+ \frac{\lambda\mu}{2\cdot 4}|Z_n^{\lambda,u}|^2\left(r_{q-1}^\lambda|Z_n^{\lambda,u}|^{2(q-1)}\right) + \frac{(\lambda\mu)^2}{2\cdot 4^2}|Z_n^{\lambda,u}|^4\left(r_{q-2}^\lambda|Z_n^{\lambda,u}|^{2(q-2)}\right)$$

$$+ \frac{(\lambda\mu)^{q/2}}{2\cdot 4^{q/2}}|Z_n^{\lambda,u}|^q\left(r_{q/2}^\lambda|Z_n^{\lambda,u}|^q\right) + (2q)!!q(q-1)2^{4q-5}(2\lambda d)^q$$

$$+ (2q)!!q(q-1)2^{4q-5}\left((2\lambda d)w_{q-1}^\lambda + (2\lambda d)^2 w_{q-2}^\lambda + (2\lambda d)^{q/2}w_{q/2}^\lambda\right)$$

$$\leq r_{q/2}^\lambda\left(\left(1 - \frac{\lambda\mu}{4}\right)^{q/2} + \frac{1}{2}\left(\frac{\lambda\mu}{4}\right)\left(1 - \frac{\lambda\mu}{4}\right)^{q/2-1}\right.$$

$$+ \frac{1}{2}\left(\frac{\lambda\mu}{4}\right)^2\left(1 - \frac{\lambda\mu}{4}\right)^{q/2-2} + \frac{1}{2}\left(\frac{\lambda\mu}{4}\right)^{q/2}\right)|Z_n^{\lambda,u}|^{2q}$$

$$+ w_q^\lambda + (2q)!!q(q-1)2^{4q-5}(2\lambda d)^q$$

$$+ (2q)!!q(q-1)2^{4q-5}\left((2\lambda d)w_{q-1}^\lambda + (2\lambda d)^2 w_{q-2}^\lambda + (2\lambda d)^{q/2}w_{q/2}^\lambda\right).$$

Using the fact that $\lambda\mu \le 1$ we get

$$
\mathbb{E}\left[|Z_{n+1}^{\lambda,u}|^{2q}|Z_n\right]
$$

$$
\le \left(1 - \frac{\lambda\mu}{2}\right)|Z_n^{\lambda,u}|^{2q} + w_q^\lambda + (2q)!!q(q-1)2^{4q-5}(2\lambda d)^q
$$

$$
+ (2q)!!q(q-1)2^{4q-5}\left((2\lambda d)w_{q-1}^\lambda + (2\lambda d)^2 w_{q-2}^\lambda + (2\lambda d)^{q/2}w_{q/2}^\lambda\right)
$$

$$
\le \left(1 - \frac{\lambda\mu}{2}\right)|Z_n^{\lambda,u}|^{2q} + (2q)!!q(q-1)2^{4q-2}(8/\mu)^{q-1}\lambda(2C)^q.
$$

Consequently, on $\{|Z_n^{\lambda,u}| \le \sqrt{8d/\mu}\{(2q)!!q(q-1)2^{4q-5}\}^{1/2}\}$, we have

$$
\mathbb{E}\left[|Z_{n+1}^{\lambda,u}|^{2q}|Z_n\right] \le \left(1 - \frac{\lambda\mu}{2}\right)|Z_n^{\lambda,u}|^{2q}
$$

$$
+ (2q)!!q(q-1)2^{4q-2}(8/\mu)^{q-1}\lambda(2C)^q
$$

$$
+ \left((2q)!!q(q-1)2^{4q-2}\right)^{q/2}(8/\mu)^{q-1}\lambda(2C)^q.
$$

Thus overall, we refine the bound in (37) to

$$
\mathbb{E}\left[|Z_{n+1}^{\lambda,u}|^{2q}\right] \le \left(1 - \frac{\lambda\mu}{2}\right)\mathbb{E}\left[|Z_n^{\lambda,u}|^{2q}\right] + M_q(8/\mu)^q C^q
$$

$$
\le C_{|z_0|,b,q,\mu}(1 + d^\theta/N + d^x)^q,
$$

where $C_{|z_0|,b,q,\mu} = \mathbb{E}\left[|Z_0^\lambda|^{2q}\right]M_q(8/\mu)^q(b+1)^q$ and $M_q = (2q)!!q(q-1)2^{6q-5}$. $\qquad\square$

Lemma 4. *Let A1, A2 and A4 hold. Then, for every $\lambda_0 < N^p/(4\mu)$, there exists a constant $C > 0$, independent of N, n, λ, d^x and d^θ, such that for any $\lambda \in (0,\lambda_0)$ one has*

$$
\mathbb{E}\left[|Z_{n+1}^{\lambda,u} - Z_n^{\lambda,u}|^4\right] \le \lambda^2 N^{-2p}C_{|z_0|,\mu,b}(1 + d^\theta/N + d^x)^2,
$$

where C is given explicitly within the proof.

Proof. We start by expanding the difference

$$
|Z_{n+1}^{\lambda,u} - Z_n^{\lambda,u}|^4 = \left(|\theta_{n+1}^{\lambda,u} - \theta_n^{\lambda,u}|^2 + \frac{1}{N}\sum_{i=1}^N |X_{n+1}^{i,\lambda,u} - X_n^{i,\lambda,u}|^2\right)^2
$$

$$
= \left(\left|\frac{-\lambda}{N^{p+1}}\sum_{i=1}^N h_{\lambda,u}^\theta(\theta_n^{\lambda,u}, X_n^{i,\lambda,u}) + \sqrt{\frac{2\lambda}{N^{p+1}}}\xi_{n+1}^{(0)}\right|^2 \right.
$$

$$
+ \frac{1}{N}\sum_{i=1}^N \left|-\frac{\lambda}{N^p}h_\lambda^x(\theta_n^{\lambda,u}, X_n^{i,\lambda,u}) + \sqrt{\frac{2\lambda}{N^p}}\xi_{n+1}^{(i)}\right|^2\Bigg)^2
$$

$$
\le \left(\frac{\lambda^2}{N^{2(p+1)}}|\sum_{i=1}^N h_{\lambda,u}^\theta(\theta_n^{\lambda,u}, X_n^{i,\lambda,u})|^2 + \frac{\lambda^2}{N^{2p+1}}\sum_{i=1}^N |h_\lambda^x(\theta_n^{\lambda,u}, X_n^{i,\lambda,u})|^2\right.
$$

$$
- 2\langle\frac{\lambda}{N^{p+1}}\sum_{i=1}^N h_{\lambda,u}^\theta(\theta_n^{\lambda,u}, X_n^{i,\lambda,u}), \sqrt{\frac{2\lambda}{N^{p+1}}}\xi_{n+1}^{(0)}\rangle
$$

$$
- 2\frac{\lambda}{N^{p+1}}\sum_{i=1}^N \langle h_{\lambda,u}^x(\theta_n^{\lambda,u}, X_n^{i,\lambda,u}), \sqrt{\frac{2\lambda}{N^p}}\xi_{n+1}^{(i)}\rangle + \frac{2\lambda}{N^{p+1}}|\xi_{n+1}^{(0)}|^2 + \frac{2\lambda}{N^{p+1}}\sum_{i=1}^N |\xi_{n+1}^{(i)}|^2\Bigg)^2
$$

Next, we further expand the expression as follows

$$
|Z_{n+1}^{\lambda,u} - Z_n^{\lambda,u}|^4 \le \frac{\lambda^4}{N^{4p+2}} \left(\sum_{i=1}^N |h_{\lambda,u}(\theta_n^{\lambda,u}, X_n^{i,\lambda,u})|^2 \right)^2 + 2 \left(\frac{\lambda^2}{N^{2p+1}} \sum_{i=1}^N |h_{\lambda,u}(\theta_n^{\lambda,u}, X_n^{i,\lambda,u})|^2 \right)
$$

$$
\times \left(-2\langle \frac{\lambda}{N^{p+1}} \sum_{i=1}^N h_{\lambda,u}^\theta(\theta_n^{\lambda,u}, X_n^{i,\lambda,u}), \sqrt{\frac{2\lambda}{N^{p+1}}} \xi_{n+1}^{(0)} \rangle \right.
$$

$$
\left. -2\frac{\lambda}{N^{p+1}} \sum_{i=1}^N \langle h_{\lambda,u}^x(\theta_n^{\lambda,u}, X_n^{i,\lambda,u}), \sqrt{\frac{2\lambda}{N^p}} \xi_{n+1}^{(i)} \rangle + \frac{2\lambda}{N^{p+1}} |\xi_{n+1}^{(0)}|^2 + \frac{2\lambda}{N^{p+1}} \sum_{i=1}^N |\xi_{n+1}^{(i)}|^2 \right)
$$

$$
+ \left| -2\langle \frac{\lambda}{N^{p+1}} \sum_{i=1}^N h_{\lambda,u}^\theta(\theta_n^{\lambda,u}, X_n^{i,\lambda,u}), \sqrt{\frac{2\lambda}{N^{p+1}}} \xi_{n+1}^{(0)} \rangle \right.
$$

$$
\left. -2\frac{\lambda}{N^{p+1}} \sum_{i=1}^N \langle h_{\lambda,u}^x(\theta_n^{\lambda,u}, X_n^{i,\lambda,u}), \sqrt{\frac{2\lambda}{N^p}} \xi_{n+1}^{(i)} \rangle + \frac{2\lambda}{N^{p+1}} |\xi_{n+1}^{(0)}|^2 + \frac{2\lambda}{N^{p+1}} \sum_{i=1}^N |\xi_{n+1}^{(i)}|^2 \right|^2. \tag{38}
$$

The first term in (38) is directly bounded by the moments of $Z_n^{\lambda,u}$,

$$
\mathbb{E}\left[\frac{\lambda^4}{N^{4p+2}} \left(\sum_{i=1}^N |h_{\lambda,u}(\theta_n^{\lambda,u}, X_n^{i,\lambda,u})|^2 \right)^2 \Big| Z_n^{\lambda,u} \right] \le \frac{\lambda^4}{N^{4p+2}} \left(2\mu^2 \sum_{i=1}^N |V_n^{i,\lambda,u}|^2 + 2\lambda^{-1}N^p \right)^2
$$

$$
\le 4\mu^4 \frac{\lambda^4}{N^{4p}} \frac{1}{N^2} \left(\sum_{i=1}^N |V_n^{i,\lambda,u}|^2 \right)^2 + 4\frac{\lambda^2}{N^{2p}} \le \mu^2 \frac{\lambda^2}{N^{2p}} |Z_n^{\lambda,u}|^4 + 4\frac{\lambda^2}{N^{2p}}. \tag{39}
$$

where the last inequality follows due to the restriction $\lambda < N^p/4\mu$. In the second term, the inner products $\langle \cdot, \cdot \rangle$ vanish under expectation, leading to

$$
2 \left(\frac{\lambda^2}{N^{2p+1}} \sum_{i=1}^N |h_{\lambda,u}(\theta_n^{\lambda,u}, X_n^{i,\lambda,u})|^2 \right) \times \left(\frac{2\lambda}{N^{p+1}} \mathbb{E}\left[|\xi_{n+1}^{(0)}|^2 | Z_n^{\lambda,u} \right] + \frac{2\lambda}{N^{p+1}} \sum_{i=1}^N \mathbb{E}\left[|\xi_{n+1}^{(i)}|^2 | Z_n^{\lambda,u} \right] \right)
$$

$$
\le 2\frac{\lambda^2}{N^{2p}} \left(2\mu^2 |Z_n^{\lambda,u}|^2 + 2\lambda^{-1}N^p \right) \left(2\frac{\lambda}{N^p} d^\theta/N + 2\frac{\lambda}{N^p} d^x \right) \le 8\frac{\lambda^3}{N^{3p}} d(\mu^2 |Z_n^{\lambda,u}|^2 + \lambda^{-1}N^p)
$$

$$
\le 2d\mu \frac{\lambda^2}{N^{2p}} |Z_n^{\lambda,u}|^2 + 8d\frac{\lambda^2}{N^{2p}}. \tag{40}
$$

By applying the usual elementary inequality and the Cauchy-Schwartz to the last term of (38), we obtain

$$
\frac{16\lambda^2}{N^{2p+2}} \left| \sum_{i=1}^N h_{\lambda,u}^\theta(\theta_n^{\lambda,u}, X_n^{i,\lambda,u}) \right|^2 \mathbb{E}\left[|\sqrt{\frac{2\lambda}{N^{p+1}}} \xi_{n+1}^{(0)}|^2 \right] + \frac{16\lambda^2}{N^{2p+1}} \sum_{i=1}^N |h_\lambda^x(\theta_n^{\lambda,u}, X_n^{i,\lambda,u})|^2 \mathbb{E}\left[|\sqrt{\frac{2\lambda}{N^p}} \xi_{n+1}^{(i)}|^2 \right]
$$

$$
+ \frac{16\lambda^2}{N^{2p+2}} \mathbb{E}\left[|\xi_{n+1}^{(0)}|^4 \right] + \frac{16\lambda^2}{N^{2p+1}} \sum_{i=1}^N \mathbb{E}\left[|\xi_{n+1}^{(i)}|^4 \right] \le 16\frac{\lambda^2}{N^{2p}} \left(2\frac{\lambda}{N^p} d^\theta/N \right) \frac{1}{N} \sum_{i=1}^N |h_{\lambda,u}^\theta(\theta_n^{\lambda,u}, X_n^{i,\lambda,u})|^2
$$

$$
+ 16\frac{\lambda^2}{N^{2p}} \left(2\frac{\lambda}{N^p} d^x \right) \frac{1}{N} \sum_{i=1}^N |h_\lambda^x(\theta_n^{\lambda,u}, X_n^{i,\lambda,u})|^2 + 16\frac{\lambda^2}{N^{2p}} (d^\theta/N)^2 + 16\frac{\lambda^2}{N^{2p}} (d^x)^2
$$

which, by using Property 1 of the taming function, can be written as

$$\frac{16\lambda^2}{N^{2p+2}}\left|\sum_{i=1}^{N}h_{\lambda,u}^{\theta}(\theta_n^{\lambda,u},X_n^{i,\lambda,u})\right|^2\mathbb{E}\left[|\sqrt{\frac{2\lambda}{N^{p+1}}}\xi_{n+1}^{(0)}|^2\right]$$

$$\leq 16\frac{\lambda^3}{N^{3p}}(2d)\frac{1}{N}\sum_{i=1}^{N}\left(2\mu^2|V_n^{i,\lambda,u}|^2+2\lambda^{-1}N^p\right)+16\frac{\lambda^2}{N^{2p}}d^2$$

$$\leq 16\cdot 4\frac{\lambda^3}{N^{3p}}\mu^2 d|Z_n^{\lambda,u}|^2+16\cdot 4\frac{\lambda^2}{N^{2p}}d+16\frac{\lambda^2}{N^{2p}}d^2$$

$$\leq 16d\mu\frac{\lambda^2}{N^{2p}}|Z_n^{\lambda,u}|^2+64\frac{\lambda^2}{N^{2p}}d+16\frac{\lambda^2}{N^{2p}}d^2. \tag{41}$$

Thus, combining (39),(40) and (41) with the bound in (38), we further refine the inequality as

$$\mathbb{E}\left[|Z_{n+1}^{\lambda,u}-Z_n^{\lambda,u}|^4|Z_n^{\lambda,u}\right]\leq\mu^2\frac{\lambda^2}{N^{2p}}|Z_n^{\lambda,u}|^4+18d\mu\frac{\lambda^2}{N^{2p}}|Z_n^{\lambda,u}|^2+4\frac{\lambda^2}{N^{2p}}(4+18d+d^2)$$

$$\leq\frac{\lambda^2}{N^{2p}}\left(\mu^2|Z_n^{\lambda,u}|^4+18d\mu|Z_n^{\lambda,u}|^2+4(4+18d+d^2)\right).$$

Finally, taking the expectation and using Lemmas 2 and 3, we conclude

$$\mathbb{E}\left[|Z_{n+1}^{\lambda,u}-Z_n^{\lambda,u}|^4\right]\leq\frac{\lambda^2}{N^{2p}}81\left(\mu^2C_{|z_0|,\mu,b}^2+18\mu C_{|z_0|,b,2,\mu}+4\right)(1+d^{\theta}/N+d^x)^2.$$

$\square$

A.2.3 Proof of Proposition 3

Proof. Consider the rescaled version (12) of the time-scaled dynamics (3)-(4) and denote by $\pi_{z,*}^N$, the rescaled invariant measure that corresponds to π_*^N, so that the rescaled initiation condition $\zeta=(\zeta_\theta,N^{-1/2}\zeta_{x^1},\ldots,N^{-1/2}\zeta_{x^N})$ follows the law $\mathcal{L}(\zeta)=\pi_{z,*}^N$. Now let $\mathcal{Z}_t^{u,N,\zeta}$ denote the aforementioned dynamics with fixed initial condition $Z_0^{u,N}=\zeta$. Then it follows that $\mathcal{L}(\mathcal{Z}_t^{u,N,\zeta})=\pi_{z,*}^N$. Since the difference $\mathcal{Z}_t^{u,N}-\mathcal{Z}_t^{u,N,\zeta}$ has vanishing diffusion, using Itô's formula and taking the expectation, one writes

$$\mathbb{E}\left[|\mathcal{Z}_t^{u,N}-\mathcal{Z}_t^{u,N,\zeta}|^2\right]=\mathbb{E}\left[|\theta_0^N-\zeta_\theta|^2\right]+\frac{1}{N}\mathbb{E}\left[|X_0^{i,N}-\zeta_{x^i}|^2\right]$$

$$-\frac{2}{N^{p+1}}\sum_{i=1}^{N}\mathbb{E}\left[\int_0^t\langle\nabla_\theta U(\mathcal{V}_s^{i,u,N})-\nabla_\theta U(\mathcal{V}_s^{i,u,N,\zeta}),\vartheta_s^{u,N}-\vartheta_s^{u,N,\zeta}\rangle ds\right]$$

$$-\frac{2}{N^{p+1}}\sum_{i=1}^{N}\mathbb{E}\left[\int_0^t\langle\nabla_x U(\mathcal{V}_s^{i,u,N})-\nabla_x U(\mathcal{V}_s^{i,u,N,\zeta}),\mathcal{X}_s^{i,u,N}-\mathcal{X}_s^{i,u,N,\zeta}\rangle ds\right]$$

$$=\mathbb{E}\left[|Z_0^N-\zeta|^2\right]$$

$$-\frac{2}{N^{p+1}}\sum_{i=1}^{N}\mathbb{E}\left[\int_0^t\langle\nabla U(\mathcal{V}_s^{i,u,N})-\nabla U(\mathcal{V}_s^{i,u,N,\zeta}),\mathcal{V}_s^{i,u,N}-\mathcal{V}_s^{i,u,N,\zeta}\rangle ds\right].$$

Then by differentiating both sides, and applying assumption A2, we obtain

$$\frac{d}{dt}\mathbb{E}\left[|\mathcal{Z}_t^{u,N}-\mathcal{Z}_t^{u,N,\zeta}|^2\right]\leq\frac{-2\mu}{N^{p+1}}\sum_{i=1}^{N}\mathbb{E}\left[|\mathcal{V}_t^{i,u,N}-\mathcal{V}_t^{i,u,N,\zeta}|^2\right]=\frac{-2\mu}{N^p}\mathbb{E}\left[|\mathcal{Z}_t^{u,N}-\mathcal{Z}_t^{u,N,\zeta}|^2\right],$$

that is

$$\frac{d}{dt}\left(\exp(2\mu/N^p)\mathbb{E}\left[|\mathcal{Z}_t^{u,N}-\mathcal{Z}_t^{u,N,\zeta}|^2\right]\right)\leq 0.$$

Integrating over $[0, t]$ and dividing by $\exp(2\mu/N^p)$ yields

$$\mathbb{E}\left[|\mathcal{Z}_t^{u,N} - \mathcal{Z}_t^{u,N,\zeta}|^2\right] \leq \exp(-2\mu/N^p)\mathbb{E}\left[|\mathcal{Z}_0^{u,N} - \mathcal{Z}_0^{u,N,\zeta}|^2\right] = \exp(-2\mu/N^p)\mathbb{E}\left[|Z_0^N - \zeta|^2\right].$$

Hence, one writes

$$\begin{aligned}
W_2\left(\mathcal{L}(\vartheta_t^{u,N}), \pi_\Theta^N\right)^2 &= \mathbb{E}\left[|\vartheta_t^{u,N} - \vartheta_t^{u,N,\zeta}|^2\right] \leq \mathbb{E}\left[|\mathcal{Z}_t^N - \mathcal{Z}_t^{N,\zeta}|^2\right] \\
&\leq \exp(-2\mu/N^p)\mathbb{E}\left[|Z_0^N - \zeta|^2\right] \\
&\leq \exp(-2\mu/N^p)\left(\mathbb{E}\left[|Z_0^N - z^*|^2\right] + \mathbb{E}\left[|z^* - \zeta|^2\right]\right) \\
&\leq \exp(-2\mu/N^p)\left(\mathbb{E}\left[|Z_0^N - z^*|^2\right] + \frac{d^x N + d^\theta}{\mu N^{p+1}}\right).
\end{aligned}$$

To bound the last term one considers the difference between $\mathcal{Z}_t^{u,N}$ and z^*, where $\nabla U(z^*) = 0$, since z^* is a minimiser. Similarly with the aforementioned calculations, one uses Itô's formula, takes the expectation and after applying assumption A2 and computing the derivative, yields

$$\frac{d}{dt}\mathbb{E}\left[|\mathcal{Z}_t^{u,N} - z^*|^2\right] \leq -2\mu\mathbb{E}\left[|\mathcal{Z}_t^{u,N} - z^*|^2\right] + 2\frac{d^x N + d^\theta}{N^{p+1}},$$

which through integration implies

$$\mathbb{E}\left[|\mathcal{Z}_t^{u,N} - z^*|^2\right] \leq \frac{d^x N + d^\theta}{\mu N^{p+1}}.$$

Therefore one immediately obtains

$$W_2(\mathcal{L}(\zeta), \delta_{z^*}) = W_2(\pi_{z,*}^N, \delta_{z^*}) \leq \limsup_{t \to \infty} W_2(\mathcal{L}(\mathcal{Z}_t^{u,N}), \delta_{z^*}) \leq \sqrt{\frac{d^x N + d^\theta}{\mu N^{p+1}}}.$$

$\square$

A.2.4 Proof of Proposition 5

Proof. We begin by considering the standard decomposition of the difference between the interpolation (27) and the scaled dynamics (28). By applying the Itô's formula to $x \to |x|^2$, we obtain

$$\begin{aligned}
|\overline{Z}_t^{\lambda,u} - \mathcal{Z}_{\lambda t}^{u,N}|^2 &= |\overline{\theta}_t^{\lambda,u} - \vartheta_{\lambda t}^{u,N}|^2 + \frac{1}{N}\sum_{i=1}^N |\overline{X}_t^{i,\lambda,u} - \mathcal{X}_{\lambda t}^{i,u,N}|^2 \\
&= -2\lambda \int_0^t \left\langle \frac{1}{N^{p+1}}\sum_{i=1}^N h_{\lambda,u}^\theta(\theta_{\lfloor t \rfloor}^{\lambda,u}, X_{\lfloor t \rfloor}^{i,\lambda,u}) - \frac{1}{N^{p+1}}\sum_{i=1}^N h^\theta(\vartheta_{\lambda t}^{u,N}, \mathcal{X}_{\lambda t}^{i,u,N}), \overline{\theta}_t^{\lambda,u} - \vartheta_{\lambda t}^{u,N}\right\rangle ds \\
&\quad -\frac{2\lambda}{N^{p+1}}\sum_{i=1}^N \int_0^t \left\langle h_{\lambda,u}^x(\theta_{\lfloor t \rfloor}^{\lambda,u}, X_{\lfloor t \rfloor}^{i,\lambda,u}) - h^x(\vartheta_{\lambda t}^{u,N}, \mathcal{X}_{\lambda t}^{i,u,N}), \overline{X}_t^{i,\lambda,u} - \mathcal{X}_{\lambda t}^{i,u,N}\right\rangle ds \\
&= -\frac{2\lambda}{N^{p+1}}\sum_{i=1}^N \int_0^t \langle h_{\lambda,u}(V_{\lfloor t \rfloor}^{i,\lambda,u}) - h(\mathcal{V}_{\lambda t}^{i,u,N}), \overline{V}_t^{i,\lambda,u} - \mathcal{V}_{\lambda t}^{i,u,N}\rangle ds.
\end{aligned}$$

To make the subsequent calculations more accessible, we introduce the following notations

$$e_t^i = \overline{V}_t^{i,\lambda,u} - \mathcal{V}_{\lambda t}^{i,u,N} \text{ and } e_t = \overline{Z}_t^{\lambda,u} - \mathcal{Z}_{\lambda t}^{u,N},$$

this implies that, $1/N \sum_{i=1}^{N} |e_t^i|^2 = |e_t|^2$. Taking the derivative of this expression, we obtain

$$\frac{d}{dt}|e_t|^2 = -\frac{2\lambda}{N^{p+1}} \sum_{i=1}^{N} \langle h_{\lambda,u}(V_{\lfloor t \rfloor}^{i,\lambda,u}) - h(\mathcal{V}_{\lambda t}^{i,u,N}), e_t^i \rangle$$

$$= -\frac{2\lambda}{N^{p+1}} \sum_{i=1}^{N} \langle h(\overline{V}_t^{i,\lambda,u}) - h(\mathcal{V}_{\lambda t}^{i,u,N}), e_t^i \rangle - \frac{2\lambda}{N^{p+1}} \sum_{i=1}^{N} \langle h_{\lambda,u}(V_{\lfloor t \rfloor}^{i,\lambda,u}) - h(V_{\lfloor t \rfloor}^{i,\lambda,u}), e_t^i \rangle$$

$$- \frac{2\lambda}{N^{p+1}} \sum_{i=1}^{N} \langle h(V_{\lfloor t \rfloor}^{i,\lambda,u}) - h(\overline{V}_t^{i,\lambda,u}), e_t^i \rangle$$

$$:= k_1(t) + k_2(t) + k_3(t). \tag{42}$$

The first term, $k_1(t)$ is controlled through the convexity of the initial potential, while the taming error $k_2(t)$ is controlled by the properties of the taming function and the moment bounds established earlier. The same bounds also control the discretisation error $k_3(t)$. Specifically we have

$$k_1(t) = -\frac{2\lambda}{N^{p+1}} \sum_{i=1}^{N} \langle h(\overline{V}_t^{i,\lambda,u}) - h(\mathcal{V}_{\lambda t}^{i,u,N}), e_t^i \rangle$$

$$\leq -\frac{2\lambda\mu}{N^{p+1}} \sum_{i=1}^{N} |e_t^i|^2 = -\frac{2\lambda\mu}{N^p}|e_t|^2. \tag{43}$$

By applying the Cauchy-Schwarz inequality and the ϵ-Young inequality with $\epsilon = \mu/2$, we can further deduce

$$k_2(t) = -\frac{2\lambda}{N^{p+1}} \sum_{i=1}^{N} \langle h_{\lambda,u}(V_{\lfloor t \rfloor}^{i,\lambda,u}) - h(V_{\lfloor t \rfloor}^{i,\lambda,u}), e_t^i \rangle$$

$$\leq \frac{2\lambda}{N^{p+1}} \sum_{i=1}^{N} \left(\frac{\mu}{4}|e_t^i|^2 + \frac{1}{\mu}|h_{\lambda,u}(V_{\lfloor t \rfloor}^{i,\lambda,u}) - h(V_{\lfloor t \rfloor}^{i,\lambda,u})|^2 \right)$$

$$\leq \frac{2\lambda\mu}{4N^p}|e_t|^2 + \frac{2\lambda}{\mu N^{p+1}} \sum_{i=1}^{N} \frac{\lambda C_1}{N^p}(1 + |V_{\lfloor t \rfloor}^{i,\lambda,u}|^{4(\ell+1)})$$

$$\leq \frac{\lambda\mu}{2N^p}|e_t|^2 + \frac{2\lambda^2 C_1}{\mu N^{2p}} + \frac{2\lambda^2 C_1}{\mu N^p} \left(\frac{1}{N} \sum_{i=1}^{N} |V_{\lfloor t \rfloor}^{i,\lambda,u}|^2 \right)^{2(\ell+1)}$$

$$\leq \frac{\lambda\mu}{2N^p}|e_t|^2 + \frac{2\lambda^2 C_1}{\mu N^p} + \frac{2\lambda^2 C_1}{\mu N^p}|Z_{\lfloor t \rfloor}^{\lambda,u}|^{4(\ell+1)}. \tag{44}$$

Note that the constant C_1 is introduced in Property 2 and p was chosen such that $p = 2\ell + 1$, ensuring that the latter expression involves a power of $|Z_{\lfloor t \rfloor}^{\lambda,u}|$. Similarly, for the final term, we derive

$$k_3(t) = -\frac{2\lambda}{N^{p+1}} \sum_{i=1}^{N} \langle h(V_{\lfloor t \rfloor}^{i,\lambda,u}) - h(\overline{V}_t^{i,\lambda,u}), e_t^i \rangle$$

$$\leq \frac{2\lambda}{N^{p+1}} \sum_{i=1}^{N} \left(\frac{\mu}{4}|e_t^i|^2 + \frac{1}{\mu}|h(V_{\lfloor t \rfloor}^{i,\lambda,u}) - h(\overline{V}_t^{i,\lambda,u})|^2 \right)$$

$$\leq \frac{\lambda\mu}{2N^p}|e_t|^2 + \frac{2\lambda C_2}{\mu N^{p+1}} \sum_{i=1}^{N} \left((1 + |V_{\lfloor t \rfloor}^{i,\lambda,u}|^{2\ell} + |\overline{V}_t^{i,\lambda,u}|^{2\ell})|V_{\lfloor t \rfloor}^{i,\lambda,u} - \overline{V}_t^{i,\lambda,u}|^2 \right)$$

$$\leq \frac{\lambda\mu}{2N^p}|e_t|^2 + \frac{2\lambda C_2}{\mu N^{p+1}} \sum_{i=1}^{N} \left(1 + |V_{\lfloor t \rfloor}^{i,\lambda,u}|^{2\ell} + |\overline{V}_t^{i,\lambda,u}|^{2\ell} \right) \sum_{i=1}^{N} |V_{\lfloor t \rfloor}^{i,\lambda,u} - \overline{V}_t^{i,\lambda,u}|^2.$$

Rearranging terms and taking expectations, and then applying the Holder's inequality, leads us to

$$
\begin{aligned}
\mathbb{E}\left[k_3(t)\right] &\leq \frac{\lambda\mu}{2N^p}\mathbb{E}\left[|e_t|^2\right] \\
&\quad + \frac{2\lambda C_2}{\mu N^{(p+1)/2}}\mathbb{E}\left[\frac{1}{N^\ell}\sum_{i=1}^N\left(1+|V_{\lfloor t\rfloor}^{i,\lambda,u}|^{2\ell}+|\overline{V}_t^{i,\lambda,u}|^{2\ell}\right)\frac{1}{N}\sum_{i=1}^N|V_{\lfloor t\rfloor}^{i,\lambda,u}-\overline{V}_t^{i,\lambda,u}|^2\right] \\
&\leq \frac{\lambda\mu}{2N^p}\mathbb{E}\left[|e_t|^2\right]+\frac{2\lambda C_2}{\mu}\mathbb{E}\left[\left(1+|Z_{\lfloor t\rfloor}^{\lambda,u}|^{2\ell}+|\overline{Z}_t^{\lambda,u}|^{2\ell}\right)|Z_{\lfloor t\rfloor}^{\lambda,u}-\overline{Z}_t^{\lambda,u}|^2\right] \\
&\leq \frac{\lambda\mu}{2N^p}\mathbb{E}\left[|e_t|^2\right]+\frac{6\lambda C_2}{\mu}\mathbb{E}^{1/2}\left[\left(1+|Z_{\lfloor t\rfloor}^{\lambda,u}|^{4\ell}+|\overline{Z}_t^{\lambda,u}|^{4\ell}\right)\right]\mathbb{E}^{1/2}\left[|Z_{\lfloor t\rfloor}^{\lambda,u}-\overline{Z}_t^{\lambda,u}|^4\right] \\
&\leq \frac{\lambda\mu}{2N^p}\mathbb{E}\left[|e_t|^2\right]+\frac{6\lambda^2 C_2 C_3}{\mu N^p}\mathbb{E}^{1/2}\left[\left(1+|Z_{\lfloor t\rfloor}^{\lambda,u}|^{4\ell}+|\overline{Z}_t^{\lambda,u}|^{4\ell}\right)\right](1+d^\theta/N+d^x) \\
&\leq \frac{\lambda\mu}{2N^p}\mathbb{E}\left[|e_t|^2\right]+\frac{6\lambda^2 C_2 C_3 C_4}{\mu N^p}(1+d^\theta/N+d^x)^{\ell+1}. \quad\quad (45)
\end{aligned}
$$

Next, taking the expectation of the remaining terms in (43) and (44) respectively and combining them with (45), we arrive, in light of (42), at

$$
\begin{aligned}
\frac{d}{dt}\mathbb{E}\left[|e_t|^2\right] &\leq -\frac{\mu\lambda}{N^p}\mathbb{E}\left[|e_t|^2\right]+\frac{2\lambda^2 C_1}{\mu N^p}+\frac{2\lambda^2 C_1 C_5}{\mu N^p}(1+d^\theta/N+d^x)^{2(\ell+1)} \\
&\quad +\frac{6\lambda^2 C_2 C_3 C_4}{\mu N^p}(1+d^\theta/N+d^x)^{\ell+1} \\
&\leq -\frac{\mu\lambda}{N^p}\mathbb{E}\left[|e_t|^2\right]+\frac{\lambda^2 C_6}{\mu N^p}(1+d^\theta/N+d^x)^{2(\ell+1)},
\end{aligned}
$$

Finally, multiplying both sides by the integrating factor $\exp\left(\lambda\mu t/N^p\right)$ and rearranging the term yields

$$
\frac{d}{dt}\left(\exp\left(\lambda\mu t/N^p\right)\mathbb{E}\left[|e_t|^2\right]\right) \leq \frac{\lambda^2 C_6}{\mu N^p}(1+d^\theta/N+d^x)^{2(\ell+1)}\exp\left(\lambda\mu t/N^p\right)
$$

$$
\mathbb{E}\left[|e_t|^2\right] \leq \lambda\frac{C_6}{\mu^2}(1+d^\theta/N+d^x)^{2(\ell+1)}.
$$

$\square$

A.3. Coordinate-wise tamed scheme

A.3.1 Key quantities for the proof of the main Lemmas.

The following definitions that refer to the rescaled dynamics (19)-(20) of the algorithm tIPLAc and its continuous time interpolations (21)-(22) are given as

$$
Z_n^{\lambda,c} = \left(\theta_{n+1}^{\lambda,c}, N^{-1/2}X_{n+1}^{1,\lambda,c},\ldots,N^{-1/2}X_{n+1}^{N,\lambda,c}\right), \quad\quad (46)
$$

$$
\overline{Z}_t^{\lambda,c} = \left(\overline{\theta}_t^{\lambda,c}, N^{-1/2}\overline{X}_t^{1,\lambda,c},\ldots,N^{-1/2}\overline{X}_t^{N,\lambda,c}\right), \quad\quad (47)
$$

$$
\mathcal{Z}_{\lambda t}^{c,N} = \left(\vartheta_{\lambda t}^{c,N}, N^{-1/2}\mathcal{X}_{\lambda t}^{1,c,N},\ldots,N^{-1/2}\mathcal{X}_{\lambda t}^{N,c,N}\right). \quad\quad (48)
$$

A.3.2 Moment and increment bounds

Lemma 5. *Let A1, A3-i and A4 hold. Then, for any $0 < \lambda < 1/(4\mu)$, it holds that,*

$$
\mathbb{E}\left[|V_n^{i,\lambda,c}|^2\right] \leq C_{|z_0|,\mu,b}(1+d^\theta+d^x),
$$

for a constant $C > 0$ independent of N, n, λ, d^x and d^θ, $\forall i \in \{1, \ldots, N\}$. If assumption A3-ii is used in place of A3-i, the lemma holds under the updated timestep constraint, $\lambda < 1/(8\mu)$.

Proof. Consider the rescaled iterates as described in equation (34), which expands as

$$
\begin{aligned}
\left|V_{n+1}^{i,\lambda,c}\right|^2 &= \left|\theta_{n+1}^{\lambda,c}\right|^2 + \left|X_{n+1}^{i,\lambda,c}\right|^2 \\
&= \left|\theta_n^{\lambda,c} - \frac{\lambda}{N}\sum_{i=1}^N h_{\lambda,c}^\theta(\theta_n^{\lambda,c}, X_n^{i,\lambda,c})\right|^2 + \frac{2\lambda}{N}|\xi_{n+1}^{(0)}|^2 \\
&\quad + 2\sqrt{\frac{2\lambda}{N}}\left\langle \theta_n^{\lambda,c} - \frac{\lambda}{N}\sum_{i=1}^N h_{\lambda,c}^\theta(\theta_n^{\lambda,c}, X_n^{i,\lambda,c}), \xi_{n+1}^{(0)}\right\rangle \\
&\quad + \left|X_n^{i,\lambda,c} - \lambda h_{\lambda,c}^x(\theta_n^{\lambda,c}, X_n^{i,\lambda,c})\right|^2 + 2\lambda|\xi_{n+1}^{(i)}|^2 \\
&\quad + 2\sqrt{2\lambda}\left\langle X_n^{i,\lambda,c} - \lambda h_{\lambda,c}^x(\theta_n^{\lambda,c}, X_n^{i,\lambda,c}), \xi_{n+1}^{(i)}\right\rangle.
\end{aligned}
$$

Taking the conditional expectation on both sides with respect to the filtration generated by $V_n^{i,\lambda,c}$, the cross terms vanish to 0, leading to

$$
\begin{aligned}
\mathbb{E}\left[\left|V_{n+1}^{i,\lambda,c}\right|^2 |V_n^{i,\lambda,c}\right] &= \mathbb{E}\left[|\theta_n^{\lambda,c}|^2 |V_n^{i,\lambda,c}\right] - \frac{2\lambda}{N}\sum_{i=1}^N \mathbb{E}\left[\langle\theta_n^{\lambda,c}, h_{\lambda,c}^\theta(\theta_n^{\lambda,c}, X_n^{i,\lambda,c})\rangle|V_n^{i,\lambda,c}\right] \\
&\quad + \frac{\lambda^2}{N^2}\mathbb{E}\left[\left|\sum_{i=1}^N h_{\lambda,c}^\theta(\theta_n^{\lambda,c}, X_n^{i,\lambda,c})\right|^2 |V_n^{i,\lambda,c}\right] + \frac{2\lambda d^\theta}{N} \\
&\quad + \mathbb{E}\left[|X_n^{i,\lambda,c}|^2 |V_n^{i,\lambda,c}\right] - 2\lambda\mathbb{E}\left[\langle X_n^{i,\lambda,c}, h_{\lambda,c}^x(\theta_n^{\lambda,c}, X_n^{i,\lambda,c})\rangle|V_n^{i,\lambda,c}\right] \\
&\quad + \lambda^2\mathbb{E}\left[|h_{\lambda,c}^x(\theta_n^{\lambda,c}, X_n^{i,\lambda,c})|^2 |V_n^{i,\lambda,c}\right] + 2\lambda d^x.
\end{aligned}
$$

Furthermore, by using the elementary inequality $(t_1 + \ldots + t_m)^p \le m^{p-1}(t_1^p + \ldots, t_m^p)$ and recognizing that all expressions within the conditional expectations are measurable, we obtain

$$
\begin{aligned}
\mathbb{E}\left[\left|V_{n+1}^{i,\lambda,c}\right|^2 |V_n^{i,\lambda,c}\right] &\le |\theta_n^{\lambda,c}| - \frac{2\lambda}{N}\sum_{i=1}^N\langle\theta_n^{\lambda,c}, h_{\lambda,c}^\theta(\theta_n^{\lambda,c}, X_n^{i,\lambda,c})\rangle + \frac{\lambda^2}{N}\sum_{i=1}^N |h_\lambda^\theta(\theta_n^{\lambda,c}, X_n^{i,\lambda,c})|^2 + \frac{2\lambda d^\theta}{N} \\
&\quad + |X_n^{i,\lambda,c}|^2 - 2\lambda\langle X_n^{i,\lambda,c}, h_{\lambda,c}^x(\theta_n^{\lambda,c}, X_n^{i,\lambda,c})\rangle + \lambda^2|h_{\lambda,c}^x(\theta_n^{\lambda,c}, X_n^{i,\lambda,c})|^2 + 2\lambda d^x \\
&\le \left|V_n^{i,\lambda,c}\right|^2 - \frac{2\lambda}{N}\sum_{i=1}^N\sum_{j=1}^{d^\theta}\theta_{n,j}^\lambda \cdot h_{\lambda,c}^{\theta_j}(\theta_n^{\lambda,c}, X_n^{i,\lambda,c}) \\
&\quad + \frac{\lambda^2}{N}\sum_{i=1}^N\sum_{j=1}^{d^\theta}|h_{\lambda,c}^{\theta_j}(\theta_n^{\lambda,c}, X_n^{i,\lambda,c})|^2 - 2\lambda\sum_{j=1}^{d^x}X_{n,j}^{i,\lambda} \cdot h_{\lambda,c}^{x_j}(\theta_n^{\lambda,c}, X_n^{i,\lambda,c}) \\
&\quad + \lambda^2\sum_{j=1}^{d^x}|h_{\lambda,c}^{x_j}(\theta_n^{\lambda,c}, X_n^{i,\lambda,c})|^2 + 2\lambda(d^\theta/N + d^x) \\
&\le \left|V_n^{i,\lambda,c}\right|^2 - \frac{2\lambda}{N}\sum_{i=1}^N\sum_{j=1}^{d^\theta}\left(\frac{\mu}{2}|\theta_{n,j}^{\lambda,c}|^2 - b\right) \\
&\quad + \frac{\lambda^2}{N}\sum_{i=1}^N\sum_{j=1}^{d^\theta}\left(2\mu^2|\theta_{n,j}^{\lambda,c}|^2 + 2\lambda^{-1}\right) - 2\lambda\sum_{j=1}^{d^x}\left(\frac{\mu}{2}|X_{n,j}^{i,\lambda,c}|^2 - b\right) \\
&\quad + \lambda^2\sum_{j=1}^{d^x}\left(2\mu^2|X_{n,j}^{i,\lambda,c}|^2 + 2\lambda^{-1}\right) + 2\lambda(d^\theta/N + d^x).
\end{aligned}
$$

Proceeding to execute the summations, we arrive at

$$\mathbb{E}\left[\left|V_{n+1}^{i,\lambda,c}\right|^2 \mid V_n^{i,\lambda,c}\right] \leq \left|V_n^{i,\lambda,c}\right|^2 - \lambda\mu|\theta_n^{\lambda,c}|^2 - \lambda\mu|X_n^{i,\lambda,c}|^2 + 2\lambda^2\mu^2|\theta_n^{\lambda,c}|^2$$
$$+ 2\lambda^2\mu^2|X_n^{i,\lambda,c}|^2 + 2\lambda b(d^\theta + d^x) + 2\lambda(d^\theta + d^x) + 2\lambda(d^\theta/N + d^x)$$
$$\leq \left(1 - \lambda\mu + 2\lambda^2\mu^2\right)\left|V_n^{i,\lambda,c}\right|^2 + 4\lambda(b+1)(d^\theta + d^x).$$

Now, consider the restriction $\lambda\mu < 1/4$, which implies

$$\mathbb{E}\left[\left|V_{n+1}^{i,\lambda,c}\right|^2\right] \leq (1-\lambda\mu/2)\mathbb{E}\left[\left|V_n^{i,\lambda,c}\right|^2\right] + 2\lambda(b+1)(d^\theta+d^x) + 2\lambda(d^\theta/N + d^x).$$

By iterating the above bound, we finally deduce

$$\mathbb{E}\left[\left|V_{n+1}^{i,\lambda,c}\right|^2\right] \leq (1-\lambda\mu/2)^n\left|V_0^{i,\lambda,c}\right|^2 + \frac{1-(1-\lambda\mu/2)^n}{\lambda\mu/2}4\lambda(b+1)(d^\theta+d^x)$$
$$\leq \left(|V_0|^2 + (8/\mu)(b+1)\right)\left(1 + d^\theta + d^x\right).$$

To establish the second part of the Lemma, we repeat the same arguments up to the point where [A3-i](#) is invoked. This leads to

$$\mathbb{E}\left[\left|V_{n+1}^{i,\lambda,c}\right|^2 \mid V_n^{i,\lambda,c}\right] \leq |V_n^{i,\lambda,c}|^2 - \frac{2\lambda}{N}\sum_{i=1}^{N}\sum_{j=1}^{d^\theta}\left(\frac{\mu}{2}|\theta_{n,j}^{\lambda,c}|^2 - \rho|X_{n,j}^{i,\lambda,c}|^2 - b\right)$$
$$+ \frac{\lambda^2}{N}\sum_{i=1}^{N}\sum_{j=1}^{d^\theta}\left(2\mu^2|\theta_{n,j}^{\lambda,c}|^2 + 2\lambda^{-1}\right)$$
$$- 2\lambda\sum_{j=1}^{d^x}\left(\frac{\mu}{2}|X_{n,j}^{i,\lambda,c}|^2 - \rho|\theta_{n,j}^{\lambda,c}|^2 - b\right)$$
$$+ \lambda^2\sum_{j=1}^{d^x}\left(2\mu^2|X_{n,j}^{i,\lambda,c}|^2 + 2\lambda^{-1}\right) + 2\lambda(d^\theta/N + d^x).$$
$$\leq \left(1 - \lambda\mu + 2\lambda\rho + 2\lambda^2\mu^2\right)\left|V_n^{i,\lambda,c}\right|^2 + 4\lambda(b+1)(d^x + d^\theta)$$

Finally, according to [A3-ii](#), since $4\rho < \mu$ implies $2\rho - \mu < -\mu/2$, and considering the constrain $\lambda\mu < 1/8$ we derive

$$\mathbb{E}\left[\left|V_{n+1}^{i,\lambda,c}\right|^2\right] \leq (1-\lambda\mu/4)\mathbb{E}\left[\left|V_n^{i,\lambda,c}\right|^2\right] + 4\lambda(b+1)(d^x + d^\theta)$$
$$\leq \left(|V_0|^2 + (16/\mu)(b+1)\right)\left(1 + d^x + d^\theta\right).$$

$\square$

Lemma 6. *Let [A1](#), [A3-i](#) and [A4](#) hold. Then, for any $0 < \lambda < 1/(4\mu)$ and $q \in [2, 2\ell+1] \cap \mathbb{N}$, it holds that,*

$$\mathbb{E}\left[|V_n^{i,\lambda,c}|^{2q}\right] \leq C_{|z_0|,\mu,b,p}(1 + d^\theta + d^x)^q,$$

for a constant $C > 0$ independent of N, n, λ, d^x and d^θ, $\forall i \in \{1, \ldots, N\}$. If assumption [A3-ii](#) is used in place of [A3-i](#), the lemma holds under the updated timestep constraint $\lambda < 1/(8\mu)$.

Proof. Similarly to the proof of Lemma [3](#), we define the following auxiliary quantities

$$A_n^\lambda := |\Delta_n^{\lambda,\theta}|^2 + |\Delta_n^{\lambda,i,x}|^2,$$
$$B_n^\lambda = 2\langle\Delta_n^\theta, G_n^{\lambda,\theta}\rangle + 2\langle\Delta_n^{\lambda,x,i}, G_n^{\lambda,x,i}\rangle + |G_n^{\lambda,\theta}|^2 + |G_n^{\lambda,x,i}|^2.$$

Now, for the 2q-th moment, we write

$$|V_{n+1}^{i,\lambda,c}|^{2q} = \left(A_n^\lambda + B_n^\lambda\right)^q$$

$$\leq (A_n^\lambda)^q + 2q(A_n^\lambda)^{q-1} B_n^\lambda + \sum_{k=2}^q \binom{q}{k} |A_n^\lambda|^{q-k} |B_n^\lambda|^k \tag{49}$$

We will handle each term separately. For the first time, we have

$$\mathbb{E}\left[|A_n^\lambda|^q | Z_n^\lambda\right] = (A_n^\lambda)^q \leq \left((1 - \lambda\mu/2)\left|V_n^{i,\lambda,c}\right|^2 + 2\lambda(b+1)(d^\theta + d^x)\right)^q$$

$$\leq \left(1 + \frac{\lambda\mu}{4}\right)^{q-1} \left(1 - \frac{\lambda\mu}{2}\right)^q |V_n^{i,\lambda,c}|^{2q} + \left(1 + \frac{4}{\lambda\mu}\right)^{q-1} 2^q \lambda^q C^q$$

$$\leq \left(1 - \frac{\lambda\mu}{4}\right)^{q-1} \left(1 - \frac{\lambda\mu}{2}\right) |V_n^{i,\lambda,c}|^{2q} + \left(\lambda + \frac{4}{\mu}\right)^{q-1} \lambda(2C)^q$$

$$\leq r_q^\lambda |V_n^{i,\lambda,c}|^{2q} + w_q^\lambda. \tag{50}$$

where $r_q^\lambda = (1 - \lambda\mu/4)^{q-1}(1 - \lambda\mu/2)$ and $w_q^\lambda = (\lambda + 4/\mu)^{q-1}\lambda(2C)^q$.
Next, we handle the second term as follows

$$\mathbb{E}\left[2q(A_n^\lambda)^{q-1} B_n^\lambda | V_n^{i,\lambda,c}\right] = 2q(A_n^\lambda)^{q-1} \mathbb{E}\left[B_n^\lambda | Z_n^\lambda\right] = 4q\lambda\left(d^\theta/N + d^x\right)(A_n^\lambda)^{q-1}$$

$$\leq 4q\lambda\left(d^\theta/N + d^x\right)\left(r_{q-1}^\lambda |V_n^{i,\lambda,c}|^{2(q-1)} + w_{q-1}^\lambda\right)$$

$$\leq 4q\,(2\lambda d)\left(r_{q-1}^\lambda |V_n^{i,\lambda,c}|^{2(q-1)} + w_{q-1}^\lambda\right). \tag{51}$$

The third term on the right-hand side of (49) can be expanded further as

$$\sum_{k=2}^q \binom{q}{k} |A_n^\lambda|^{q-k} |B_n^\lambda|^k = \sum_{\ell=0}^{q-2} \binom{q}{\ell+2} |A_n^\lambda|^{q-2-k} |B_n^\lambda|^{k+2}$$

$$= \frac{q}{\ell+2}\frac{q-1}{\ell+1} \sum_{\ell=0}^{q-2} \binom{q}{\ell} |A_n^\lambda|^{q-2-k} |B_n^\lambda|^k |B_n^\lambda|^2$$

$$\leq q(q-1)\left(|A_n^\lambda| + |B_n^\lambda|\right)^{q-2} |B_n^\lambda|^2$$

$$\leq q(q-1)2^{q-3}|A_n^\lambda|^{q-2}|B_n^\lambda|^2 + q(q-1)2^{q-3}|B_n^\lambda|^q.$$

$$= D + F \tag{52}$$

Taking the expectation of the above terms, we get

$$\mathbb{E}[D | V_n^{i,\lambda,c}] = q(q-1)2^{q-3}|A_n^\lambda|^{q-2}\mathbb{E}\left[|B_n^\lambda|^2 | V_n^{i,\lambda,c}\right],$$

where

$$\mathbb{E}\left[|B_n^\lambda|^2 | V_n^{i,\lambda,c}\right]$$

$$= \mathbb{E}\left[\left|2\langle \Delta_n^\theta, G_n^{\lambda,\theta}\rangle + 2\langle \Delta_n^{\lambda,x,i}, G_n^{\lambda,x,i}\rangle + |G_n^{\lambda,\theta}|^2 + |G_n^{\lambda,x,i}|^2\right|^2 | V_n^{i,\lambda,c}\right]$$

$$\leq 4\left(\mathbb{E}\left[4|\Delta_n^{\lambda,\theta}|^2|G_n^{\lambda,\theta}|^2|V_n^{i,\lambda,c}\right] + \mathbb{E}\left[4|\Delta_n^{\lambda,x,i}|^2|G_n^{\lambda,x,i}|^2|V_n^{i,\lambda,c}\right]\right.$$

$$\left. + \mathbb{E}\left[|G_n^{\lambda,x,i}|^4|V_n^{i,\lambda,c}\right] + \mathbb{E}\left[|G_n^{\lambda,x,i}|^4|V_n^{i,\lambda,c}\right]\right)$$

$$\leq 4\left(4(2\lambda d^\theta/N)|\Delta_n^{\lambda,\theta}|^2 + 4(2\lambda d^x)|\Delta_n^{\lambda,x,i}|^2 + 12(2\lambda d^\theta/N)^2 + 12(2\lambda d^x)^2\right)$$

$$\leq 16(2\lambda d)\left(|\Delta_n^{\lambda,\theta}|^2 + |\Delta_n^{\lambda,x,i}|^2\right) + 96(2\lambda d)^2$$

$$\leq 16(2\lambda d)A_n^\lambda + 96(2\lambda d)^2.$$

with $d = \max\{d^\theta/N, d^x\}$. Substituting this result for the term D, we derive

$$\mathbb{E}[D|V_n^{i,\lambda,c}] = q(q-1)2^{q-3}|A_n^\lambda|^{q-2}\left(16(2\lambda d)A_n^\lambda + 96(2\lambda d)^2\right)$$
$$= 16(2\lambda d)q(q-1)2^{q-3}\left(r_{q-1}^\lambda|V_n^\lambda|^{2(q-1)} + w_{q-1}^\lambda\right)$$
$$+ 96(2\lambda d)^2 q(q-1)2^{q-3}\left(r_{q-2}^\lambda|V_n^\lambda|^{2(q-2)} + w_{q-2}^\lambda\right).$$

Additionally

$$\mathbb{E}[F|V_n^{i,\lambda,c}] = q(q-1)2^{q-3}|B_n^\lambda|^q$$
$$= q(q-1)2^{q-3}\mathbb{E}\left[\left|2\langle\Delta_n^\theta, G_n^{\lambda,\theta}\rangle + 2\langle\Delta_n^{\lambda,x,i}, G_n^{\lambda,x,i}\rangle + |G_n^{\lambda,\theta}|^2 + |G_n^{\lambda,x,i}|^2\right|^q|V_n^{i,\lambda,c}\right]$$
$$\le q(q-1)2^{q-3}4^{q-1}\left(\mathbb{E}\left[2^q|\Delta_n^{\lambda,\theta}|^q|G_n^{\lambda,\theta}|^q|V_n^{i,\lambda,c}\right] + \mathbb{E}\left[2^q|\Delta_n^{\lambda,x,i}|^q|G_n^{\lambda,x,i}|^q|V_n^{i,\lambda,c}\right]\right.$$
$$\left.+\mathbb{E}\left[|G_n^{\lambda,\theta}|^{2q}|V_n^{i,\lambda,c}\right] + \mathbb{E}\left[|G_n^{\lambda,x,i}|^{2q}|V_n^{i,\lambda,c}\right]\right)$$
$$\le q(q-1)2^{3q-5}\left(2^q(2\lambda d^\theta/N)^{q/2}q!!\left(|\Delta_n^{\lambda,\theta}|^2\right)^{q/2} + 2^q q!!(2\lambda d^x)^{q/2}\left(|\Delta_n^{\lambda,x,i}|^2\right)^{q/2}\right.$$
$$\left.+(2q)!!(2\lambda d^\theta/N)^q + (2q)!!(2\lambda d^x)^q\right)$$
$$\le (q(q-1))^{q+1}2^{4q-5}\left((2\lambda d)^{q/2}\left(|\Delta_n^{\lambda,\theta}|^2 + |\Delta_n^{\lambda,x,i}|^2\right)^{q/2} + 2(2\lambda d)^q\right)$$
$$\le (q(q-1))^{q+1}2^{4q-5}\left((2\lambda d)^{q/2}\left|A_n^\lambda\right|^{q/2} + 2(2\lambda d)^q\right)$$
$$\le (q(q-1))^{q+1}2^{4q-5}(2\lambda d)^{q/2}\left(r_{q/2}^\lambda|V_n^\lambda|^q + w_{q/2}^\lambda\right)$$
$$+ (q(q-1))^{q+1}2^{4q-5}2(2\lambda d)^q.$$

Combining the above results in view of (52), yields

$$\mathbb{E}\left[\sum_{k=2}^q \binom{q}{k}|A_n^\lambda|^{q-k}|B_n^\lambda|^k\right]$$
$$\le 16(2\lambda d)q(q-1)2^{q-3}\left(r_{q-1}^\lambda|V_n^\lambda|^{2(q-1)} + w_{q-1}^\lambda\right)$$
$$+ 96(2\lambda d)^2 q(q-1)2^{q-3}\left(r_{q-2}^\lambda|V_n^\lambda|^{2(q-2)} + w_{q-2}^\lambda\right)$$
$$+ (q(q-1))^{q+1}2^{4q-5}(2\lambda d)^{q/2}\left(r_{q/2}^\lambda|V_n^\lambda|^q + w_{q/2}^\lambda\right)$$
$$+ (q(q-1))^{q+1}2^{4q-4}(2\lambda d)^q. \tag{53}$$

Substituting (50),(51) and (53) into (49) gives the following

$$\mathbb{E}\left[|V_{n+1}^{i,\lambda,c}|^{2q}|V_n^{i,\lambda,c}\right] \le r_q^\lambda|V_n^{i,\lambda,c}|^{2q} + w_q^\lambda$$
$$+ 4q(2\lambda d)\left(1 + 4(q-1)2^{q-3}\right)\left(r_{q-1}^\lambda|V_n^{i,\lambda,c}|^{2(q-1)} + w_{q-1}^\lambda\right)$$
$$+ 96(2\lambda d)^2 q(q-1)2^{q-3}\left(r_{q-2}^\lambda|V_n^{i,\lambda,c}|^{2(q-2)} + w_{q-2}^\lambda\right)$$
$$+ (q(q-1))^{q+1}2^{4q-5}(2\lambda d)^{q/2}\left(r_{q/2}^\lambda|V_n^{i,\lambda,c}|^q + w_{q/2}^\lambda\right)$$
$$+ (q(q-1))^{q+1}2^{4q-4}(2\lambda d)^q. \tag{54}$$

Now, consider

$$|V_n^{i,\lambda,c}| \ge \sqrt{8d/\mu}\left\{2(q(q-1))^{q+1}2^{4(q-1)}\right\}^{1/2} \ge \sqrt{8d/\mu}\left\{2(q(q-1))^{q+1}2^{4(q-1)}\right\}^{1/q}.$$

From this, we observe

$$\mathbb{E}\left[|V_{n+1}^{i,\lambda,c}|^{2q}|V_n^{i,\lambda,c}\right]$$

$$\leq r_q^\lambda |V_n^{i,\lambda,c}|^{2q} + w_q^\lambda + \frac{\lambda\mu}{2\cdot 4}|V_n^{i,\lambda,c}|^2\left(r_{q-1}^\lambda |V_n^{i,\lambda,c}|^{2(q-1)}\right)$$

$$+ \frac{(\lambda\mu)^2}{2\cdot 4^2}|V_n^{i,\lambda,c}|^4\left(r_{q-2}^\lambda |V_n^{i,\lambda,c}|^{2(q-2)}\right)$$

$$+ \frac{(\lambda\mu)^{q/2}}{2\cdot 4^{q/2}}|V_n^{i,\lambda,c}|^q\left(r_{q/2}^\lambda |V_n^{i,\lambda,c}|^q\right) + (q(q-1))^{q+1}2^{4(q-1)}(2\lambda d)^q$$

$$+ (q(q-1))^{q+1}2^{4(q-1)}\left((2\lambda d)w_{q-1}^\lambda + (2\lambda d)^2 w_{q-2}^\lambda + (2\lambda d)^{q/2}w_{q/2}^\lambda\right)$$

$$\leq r_{q/2}^\lambda\left(\left(1-\frac{\lambda\mu}{4}\right)^{q/2} + \frac{1}{2}\left(\frac{\lambda\mu}{4}\right)\left(1-\frac{\lambda\mu}{4}\right)^{q/2-1}\right.$$

$$+\frac{1}{2}\left(\frac{\lambda\mu}{4}\right)^2\left(1-\frac{\lambda\mu}{4}\right)^{q/2-2} + \left.\frac{1}{2}\left(\frac{\lambda\mu}{4}\right)^{q/2}\right)|V_n^{i,\lambda,c}|^{2q}$$

$$+ w_q^\lambda + (q(q-1))^{q+1}2^{4(q-1)}(2\lambda d)^q$$

$$+ (q(q-1))^{q+1}2^{4(q-1)}\left((2\lambda d)w_{q-1}^\lambda + (2\lambda d)^2 w_{q-2}^\lambda + (2\lambda d)^{q/2}w_{q/2}^\lambda\right).$$

Using the fact that $\lambda\mu \leq 1$, we further simplify the bound

$$\mathbb{E}\left[|V_{n+1}^{i,\lambda,c}|^{2q}|V_n^{i,\lambda,c}\right]$$

$$\leq \left(1-\frac{\lambda\mu}{2}\right)|V_n^{i,\lambda,c}|^{2q}$$

$$+ w_q^\lambda + (q(q-1))^{q+1}2^{4(q-1)}(2\lambda d)^q$$

$$+ (q(q-1))^{q+1}2^{4(q-1)}\left((2\lambda d)w_{q-1}^\lambda + (2\lambda d)^2 w_{q-2}^\lambda + (2\lambda d)^{q/2}w_{q/2}^\lambda\right)$$

$$\leq \left(1-\frac{\lambda\mu}{2}\right)|V_n^{i,\lambda,c}|^{2q} + (q(q-1))^{q+1}2^{4(q-1)}10^q C^q(d\lambda)^q.$$

Consequently, on $\{|V_n^{i,\lambda,c}| \leq \sqrt{8d/\mu}\left\{2(q(q-1))^{q+1}2^{4(q-1)}\right\}^{1/2}\}$, we derive

$$\mathbb{E}\left[|V_{n+1}^{i,\lambda,c}|^{2q}|V_n^{i,\lambda,c}\right] \leq \left(1-\frac{\lambda\mu}{2}\right)|V_n^{i,\lambda,c}|^{2q}$$

$$+ (q(q-1))^{q+1}2^{4(q-1)}10^q C^q(d\lambda)^q$$

$$+ 3\left(q(q-1)^{q+1}2^{4(q-1)}\right)^q(2d)^q\left(\frac{4}{\mu}\right)^q 2^{q-1}.$$

Thus, the bound provided in (54) is improved to

$$\mathbb{E}\left[|V_{n+1}^{i,\lambda,c}|^{2q}\right] \leq \left(1-\frac{\lambda\mu}{2}\right)\mathbb{E}\left[|V_n^{i,\lambda,c}|^{2q}\right] + C(d,q,\lambda,\mu),$$

Hence, the usual bound for recursive sequences follows, completing the proof. $\qquad\square$

A.3.3 Proof of Proposition 6

Proof. We begin by considering the standard decomposition of the difference between the interpolation (35) and the scaled dynamics (36). By applying the Itô's formula to $x \to |x|^2$, we obtain

$$|\overline{Z}_t^{\lambda,c} - \mathcal{Z}_{\lambda t}^{c,N}|^2 = |\overline{\theta}_t^{\lambda,c} - \vartheta_{\lambda t}^{c,N}|^2 + \frac{1}{N}\sum_{i=1}^N |\overline{X}_t^{i,\lambda,c} - \mathcal{X}_{\lambda t}^{i,c,N}|^2$$

$$= -2\lambda \int_0^t \left\langle \frac{1}{N}\sum_{i=1}^N h_{\lambda,c}^\theta(\theta_{\lfloor t\rfloor}^{\lambda,c}, X_{\lfloor t\rfloor}^{i,\lambda,c}) - \frac{1}{N}\sum_{i=1}^N h^\theta(\vartheta_{\lambda t}^{c,N}, \mathcal{X}_{\lambda t}^{i,c,N}), \overline{\theta}_t^{i,\lambda,c} - \vartheta_{\lambda t}^{c,N} \right\rangle ds$$

$$- \frac{2\lambda}{N}\sum_{i=1}^N \int_0^t \left\langle h_{\lambda,c}^x(\theta_{\lfloor t\rfloor}^{\lambda,c}, X_{\lfloor t\rfloor}^{i,\lambda,c}) - h^x(\vartheta_{\lambda t}^{c,N}, \mathcal{X}_{\lambda t}^{i,c,N}), \overline{X}_t^{i,\lambda,c} - \mathcal{X}_{\lambda t}^{i,c,N} \right\rangle ds$$

$$= -\frac{2\lambda}{N}\sum_{i=1}^N \int_0^t \langle h_{\lambda,c}(V_{\lfloor t\rfloor}^{i,\lambda,c}) - h(\mathcal{V}_{\lambda t}^{i,c,N}), \overline{V}_t^{i,\lambda,c} - \mathcal{V}_{\lambda t}^{i,c,N} \rangle ds.$$

In order to make the forthcoming calculations more readable, we define the following quantities

$$e_t^i = \overline{V}_t^{i,\lambda,c} - \mathcal{V}_{\lambda t}^{i,c,N} \quad \text{and} \quad e_t = \overline{Z}_t^{\lambda,c} - \mathcal{Z}_{\lambda t}^{c,N}.$$

Hence it follows that $1/N \sum_{i=1}^N |e_t^i|^2 = |e_t|^2$. Taking the derivative of this expression, we arrive at

$$\frac{d}{dt}|e_t|^2 = -\frac{2\lambda}{N}\sum_{i=1}^N \langle h_{\lambda,c}(V_{\lfloor t\rfloor}^{i,\lambda,c}) - h(\mathcal{V}_{\lambda t}^{i,c,N}), e_t^i \rangle$$

$$= -\frac{2\lambda}{N}\sum_{i=1}^N \langle h(\overline{V}_t^{i,\lambda,c}) - h(\mathcal{V}_{\lambda t}^{i,c,N}), e_t^i \rangle$$

$$- \frac{2\lambda}{N}\sum_{i=1}^N \langle h_{\lambda,c}(V_{\lfloor t\rfloor}^{i,\lambda,c}) - h(V_{\lfloor t\rfloor}^{i,\lambda,c}), e_t^i \rangle$$

$$- \frac{2\lambda}{N}\sum_{i=1}^N \langle h(V_{\lfloor t\rfloor}^{i,\lambda,c}) - h(\overline{V}_t^{i,\lambda,c}), e_t^i \rangle$$

$$= k_1(t) + k_2(t) + k_3(t). \tag{55}$$

The first term $k_1(t)$ is controlled through the convexity of the initial potential. The taming error $k_2(t)$ is controlled by the properties of the taming function and the additional moment bounds we have already established. These bounds also allow us to control the discretisation error $k_3(t)$. Specifically, we have

$$k_1(t) = -\frac{2\lambda}{N}\sum_{i=1}^N \langle h(\overline{V}_t^{i,\lambda,c}) - h(\mathcal{V}_{\lambda t}^{i,c,N}), e_t^i \rangle \le -\frac{2\lambda\mu}{N}\sum_{i=1}^N |e_t^i|^2 = -2\lambda\mu|e_t|^2. \tag{56}$$

By applying the Cauchy-Schwarz inequality and the ϵ-Young inequality with $\epsilon = \mu/2$, we get

$$k_2(t) = -\frac{2\lambda}{N}\sum_{i=1}^N \langle h_{\lambda,c}(V_{\lfloor t\rfloor}^{i,\lambda,c}) - h(V_{\lfloor t\rfloor}^{i,\lambda,c}), e_t^i \rangle \le \frac{2\lambda}{N}\sum_{i=1}^N \left(\frac{1}{\mu}|h_{\lambda,c}(V_{\lfloor t\rfloor}^{i,\lambda,c}) - h(V_{\lfloor t\rfloor}^{i,\lambda,c})|^2 + \frac{\mu}{4}|e_t^i|^2 \right)$$

$$\le 2\frac{\lambda\mu}{4}|e_t|^2 + \frac{2\lambda}{\mu N}\sum_{i=1}^N 2\lambda C_2^2(1 + |V_{\lfloor t\rfloor}^{i,\lambda,c}|^{4(\ell+1)}) \le 2\frac{\lambda\mu}{4}|e_t|^2 + \frac{4\lambda^2}{\mu}C_2^2 + \frac{4\lambda^2}{\mu N}\sum_{i=1}^N |V_{\lfloor t\rfloor}^{i,\lambda,c}|^{4(\ell+1)}. \tag{57}$$

Similarly, for the final term, we derive

$$k_3(t) = -\frac{2\lambda}{N}\sum_{i=1}^N \langle h(V_{\lfloor t \rfloor}^{i,\lambda,c}) - h(\overline{V}_t^{i,\lambda,c}), e_t^i\rangle \le \frac{2\lambda}{N}\sum_{i=1}^N\left(\frac{1}{\mu}|h(V_{\lfloor t \rfloor}^{i,\lambda,c}) - h(\overline{V}_t^{i,\lambda,c})|^2 + \frac{\mu}{4}|e_t^i|^2\right)$$

$$\le 2\frac{\lambda\mu}{4}|e_t|^2 + \frac{2\lambda L^2}{\mu N}\sum_{i=1}^N\left((1 + |V_{\lfloor t \rfloor}^{i,\lambda,c}|^\ell + |\overline{V}_t^{i,\lambda,c}|^\ell)^2|V_{\lfloor t \rfloor}^{i,\lambda,c} - \overline{V}_t^{i,\lambda,c}|^2\right). \tag{58}$$

Now, combining the bounds in (56),(57) and (58) and substituting them in (55), we can take the expectation, which leads to

$$\frac{d}{dt}\mathbb{E}\left[|e_t|^2\right] \le -\mu\lambda\mathbb{E}\left[|e_t|^2\right] + \frac{4\lambda^2 C_2^2}{\mu} + \frac{4\lambda^2}{\mu}\frac{1}{N}\sum_{i=1}^N\mathbb{E}\left[|V_{\lfloor t \rfloor}^{i,\lambda,c}|^{4(\ell+1)}\right]$$

$$+ \frac{2\lambda L^2}{\mu}\frac{1}{N}\sum_{i=1}^N\mathbb{E}^{1/2}\left[(1 + |V_{\lfloor t \rfloor}^{i,\lambda,c}|^\ell + |\overline{V}_t^{i,\lambda,c}|^\ell)^4\right]\mathbb{E}^{1/2}\left[|V_{\lfloor t \rfloor}^{i,\lambda,c} - \overline{V}_t^{i,\lambda,c}|^4\right]$$

$$\le -\mu\lambda\mathbb{E}\left[|e_t|^2\right] + \frac{4\lambda^2(C_2^2 + C_3)}{\mu}$$

$$+ \frac{2\lambda L^2}{\mu}\left(6\max\{1, 2C_4\}\right)\lambda\left(\mu C_5 + 6\sqrt{d\mu}C_6 + 10d + 2\right)$$

$$\le -\mu\lambda\mathbb{E}\left[|e_t|^2\right] + \lambda^2 M(L,\mu,d,\ell) \le \lambda\left(-\frac{\mu}{2}\mathbb{E}\left[|e_t|^2\right] + \lambda M\right).$$

Finally, multiplying both sides by the integrating factor $\exp\left(\lambda\mu t/2\right)$ and rearranging the terms, we conclude

$$\frac{d}{dt}\left(\exp\left(\lambda\mu t/2\right)\mathbb{E}\left[|e_t|^2\right]\right) \le \lambda^2 M\exp\left(\lambda\mu t/2\right)$$

$$\mathbb{E}\left[|e_t|^2\right] \le \lambda\frac{2M}{\mu}.$$

$\square$

B. Verifying Assumptions in Model Examples

B.1. Bayesian logistic regression with thin tails

Remark 3. *Consider the potential $U(\theta, x)$ as given in (4.2), it satisfies A1, i.e. it is locally Lipschitz with polynomial growth.*

Proof. It is quite straightforward to compute

$$|\nabla U(v) - \nabla U(v')| \le |\nabla U_x(v) - \nabla U_x(v')| + |\nabla U_\theta(v) - \nabla U_\theta(v')|$$

$$\le \frac{8}{\sigma^4}\left(x - \theta - x' + \theta'\right)\left(3/2|x - \theta|^2 + 3/2|x' - \theta'|^2\right)$$

$$+ \frac{4}{\sigma^2}\left(|x - x'| + |\theta - \theta'|\right) + 1/4\sum_{j=1}^{d^y}|u_j|^2|x - x'|$$

$$\le \frac{16}{\sigma^4}\left(|v - v'|\right)\left(4|v|^2 + 4|v'|^2\right) + \left(\frac{8}{\sigma^2} + 1/4\sum_{j=1}^{d^y}|u_j|^2\right)|v - v'|$$

$$\le \max\left(\frac{64}{\sigma^4}, \frac{8}{\sigma^2} + 1/4\sum_{j=1}^{d^y}|u_j|^2\right)\left(1 + |v|^2 + |v'|^2\right)|v - v'|.$$

We note that the last term in the second inequality follows from the fact that the logistic function is Lipschitz continuous with constant $1/4$. $\square$

Remark 4. *Consider the potential $U(\theta, x)$ as given in (4.2), it satisfies A2 with $\mu = 2/\sigma^2$, i.e. it is strongly convex.*

Proof. First, we compute the second-order partial derivatives as follows

$$\frac{\partial^2 U_y}{\partial \theta_i \theta_r} = \delta_{ir} \left(\frac{12}{\sigma^4}(x_i - \theta_i)^2 + \frac{2}{\sigma^2} \right),$$

$$\frac{\partial^2 U_y}{\partial x_i x_r} = \delta_{ir} \left(\frac{12}{\sigma^4}(x_i - \theta_i)^2 + \frac{2}{\sigma^2} \right) + \sum_{j=1}^{dy} s(u_j^T x) \left(1 - s(u_j^T x) \right) u_{j,i} u_{j,r},$$

$$\frac{\partial^2 U_y}{\partial x_i \theta_r} = -\delta_{ir} \left(\frac{12}{\sigma^4}(x_i - \theta_i)^2 + \frac{2}{\sigma^2} \right).$$

We define the diagonal matrix A, with elements $a_{ir} = \delta_{ir} \left(12/\sigma^4 (x_i - \theta_i)^2 + 2/\sigma^2 \right)$, to write the Hessian of $U(\theta, x)$ in the following form

$$H\left(U(\theta, x) \right) = \begin{bmatrix} A & -A \\ -A & A \end{bmatrix} + \left[0, 0; 0, \sum_{j=1}^{dy} s(u_j^T x) \left(1 - s(u_j^T x) \right) u_j \otimes u_j \right].$$

From this we observe that the leftmost term is a positive definite matrix due to its diagonal block structure, while the second term is semi-positive definite (see further details in [10]). Hence $H(U(\theta, x))$ is positive definite, making $U(\theta, x)$ strongly convex with a convexity constant $\mu = 2/\sigma^2$. □

B.2. Higher order - Toy Example

Remark 5. *Consider the potential $U(\theta, x) = |x|^{4m} + (|x|^{2m} + 1)(|\theta|^{2m} + 1) + |\theta|^{4m}$. Then $U(\theta, x)$ satisfies A1, i.e. it is locally Lipschitz with polynomial growth and constant $L = 2m2^{4m}$.*

Proof. We decompose the difference of the gradient of U as follows

$$|\nabla U(v) - \nabla U(v')| \leq |\nabla_x U(\theta, x) - \nabla_x U(x', \theta')| + |\nabla_\theta U(\theta, x) - \nabla_\theta U(x', \theta')| \tag{59}$$

Due to the symmetry of U it suffices to consider only the first term incorporating $\nabla_x U$,

$$\begin{aligned}
&|\nabla_x U(\theta, x) - \nabla_x U(x', \theta')| \\
&= \big| 4m|x|^{4m-2}x + 2m \left(|\theta|^{2m} + 1 \right) |x|^{2m-2}x \\
&\quad -4m|x'|^{4m-2}x' - 2m \left(|\theta'|^{2m} + 1 \right) |x'|^{2m-2}x' \big| \\
&\leq 4m \left| |x|^{4m-2}x - |x'|^{4m-2}x' \right| + 2m \big| \left(|\theta|^{2m} + 1 \right) |x|^{2m-2}x \\
&\quad - \left(|\theta'|^{2m} + 1 \right) |x'|^{2m-2}x' \big| \\
&\leq 4m|x| \left| |x|^{4m-2} - |x'|^{4m-2} \right| + 4m|x'|^{4m-2}|x - x'| \\
&\quad + 2m \left(|\theta|^{2m} + 1 \right) |x|^{2m-2}|x - x'| \\
&\quad + 2m|x'| \left| \left(|\theta|^{2m} + 1 \right) |x|^{2m-2} - \left(|\theta'|^{2m} + 1 \right) |x'|^{2m-2} \right| \\
&\leq 4m \sum_{i=0}^{4m-3} |x|^{4m-2-i}|x'|^i |x - x'| + 4m|x'|^{4m-2}|x - x'| \\
&\quad + 2m|x|^{2m-2}|x - x'| + 2m|v|^{4m-2}|x - x'| \\
&\quad + 2m|x'| \left| |x|^{2m-2} - |x'|^{2m-2} \right| + 2m|x'||\theta|^{2m} \left| |x|^{2m-2} - |x'|^{2m-2} \right| \\
&\quad + 2m|x'|^{2m-1} \left| |\theta|^{2m} - |\theta'|^{2m} \right|.
\end{aligned} \tag{60}$$

Next, we define the expression J_x which contains all the terms with common factor $|x - x'|$. Notice that due to the symmetry of the problem, the second term of (59) yields only one such term, corresponding to the last term in (60), that is

$2m|\theta'|^{2m-1}\left||x|^{2m}-|x'|^{2m}\right|$. Hence J_x is bounded as follows

$$
\begin{aligned}
J_x = 2m & \left(2 \sum_{i=0}^{4m-2} |x|^{4m-2-i}|x'|^i + |v|^{4m-2} + \sum_{i=0}^{2m-2} |x|^{2m-2-i}|x'|^i \right. \\
& \left. + |\theta|^{2m} \sum_{i=0}^{2m-3} |x|^{2m-3-i}|x'|^{i+1} + |\theta'|^{2m-1} \sum_{i=0}^{2m-1} |x|^{2m-1-i}|x'|^i \right) |x-x'| \\
\le 2m & \left(3 \sum_{i=0}^{4m-2} |x|^{4m-2-i}|x'|^i + |v|^{4m-2} + \frac{m(2m-3)}{2m-1}|\theta|^{4m-2} \right. \\
& + \frac{m-1}{2m-1} \sum_{i=0}^{2m-3} |x|^{(2m-3-i)(2m-1/m-1)}|x'|^{(i+1)(2m-1/m-1)} \\
& \left. + \frac{2m-1}{2}|\theta'|^{4m-2} + \frac{1}{2} \sum_{i=0}^{4m-2} |x|^{4m-2-i}|x'|^i \right) |x-x'| \\
\le 2m & \left(\frac{3}{2} \sum_{i=0}^{4m-2} |x|^{4m-2-i}|x'|^i + |v|^{4m-2} + (2m-3)|\theta|^{4m-2} + \frac{1}{2}|\theta'|^{4m-2} \right. \\
& \left. + \frac{1}{2} \left(|x|^{2m-1/m-1} + |x'|^{2m-1/m-1} \right)^{2m-2} \right) |x-x'| \\
\le 2m & \left(\left(3 \cdot 2^{4m-4} + 2^{2m-4} \right) \left(|x|^{4m-2} + |x'|^{4m-2} \right) \right. \\
& \left. + |v|^{4m-2} + (2m-3)|\theta|^{4m-2} + \frac{1}{2}|\theta'|^{4m-2} \right) |x-x'| \\
\le 2m & \left(3 \cdot 2^{4m-4} + 2^{2m-4} + 1 \right) \left(|v|^{4m-2} + |v'|^{4m-2} \right) |v-v'|. \tag{61}
\end{aligned}
$$

The same bound as in (61) can be derived for J_θ, by following the above calculations. Thus, we obtain the following bound for (59)

$$
\begin{aligned}
|\nabla U(v) - \nabla U(v')| &\le J_x + J_\theta \\
&\le \frac{m}{4} \left(3 \cdot 2^{4m} + 2^{2m} + 16 \right) \left(1 + |v|^{4m-2} + |v'|^{4m-2} \right) |v-v'| \\
&\le 2m2^{4m} \left(1 + |v|^{4m-2} + |v'|^{4m-2} \right) |v-v'|. \tag{62}
\end{aligned}
$$

$\square$

Remark 6. *Consider the potential $U(\theta, x)$ as given in (25), it satisfies A2, i.e. it is strongly convex with $\mu = 2$.*

Proof. First, we compute the second-order partial derivatives as follows

$$
\begin{aligned}
\frac{\partial^2 U}{\partial x_i \partial x_j} &= \left(4m|x|^{4m-2} + 2m(|\theta|^{2m}+1)|x|^{2m-2} + 4|x|^2 + 2(|\theta|^2+1) \right) \delta_{ij} \\
&\quad + \left(8m(2m-1)|x|^{4m-4} + 4m(m-1)(|\theta|^{2m}+1)|x|^{2m-4} + 8 \right) x_i x_j \\
&:= A(x,\theta)\delta_{ij} + B(x,\theta)x_i x_j + 8x_i\theta_j \\
\frac{\partial^2 U}{\partial x_i \partial \theta_j} &= (4m^2|x|^{2m-2}|\theta|^{2m-2} + 4)x_i\theta_j := C(x,\theta)x_i\theta_j + 4x_i\theta_j \\
\frac{\partial^2 U}{\partial \theta_i \partial \theta_j} &= \left(4m|\theta|^{4m-2} + 2m(|x|^{2m}+1)|\theta|^{2m-2} + 4|\theta|^2 + 2(|x|^2+1) \right) \delta_{ij} \\
&\quad + \left(8m(2m-1)|\theta|^{4m-4} + 4m(m-1)(|x|^{2m}+1)|\theta|^{2m-4} + 8 \right) \theta_i\theta_j \\
&= A(\theta,x)\delta_{ij} + B(\theta,x)\theta_i\theta_j + 8x_i\theta_j
\end{aligned}
$$

Next, we decompose the Hessian of $U(\theta, x)$ in the following way

$$H(U(\theta, x))) = \begin{bmatrix} A(x,\theta)I_{d^x} & 0 \\ 0 & A(\theta,x)I_{d^\theta} \end{bmatrix} + \begin{bmatrix} B(x,\theta)J_{d^x \times d^x} & C(x,\theta)J_{d^x \times d^\theta} \\ C(x,\theta)J_{d^\theta \times d^x} & B(\theta,x)J_{d^\theta \times d^\theta} \end{bmatrix} \odot \begin{bmatrix} x \\ \theta \end{bmatrix} \begin{bmatrix} x & \theta \end{bmatrix}$$
$$+ \begin{bmatrix} 8J_{d^x \times d^x} & 4J_{d^x \times d^\theta} \\ 4J_{d^\theta \times d^x} & 8J_{d^\theta \times d^\theta} \end{bmatrix} \odot \begin{bmatrix} x \\ \theta \end{bmatrix} \begin{bmatrix} x & \theta \end{bmatrix},$$

where $\odot$ denotes the Hadamard product and $J_{n,m}$ the $n \times m$ ones-matrix. One observes that the leftmost matrix is positive definite with constant $\mu = 2$ due to its positive diagonal structure, while the second term is semi-positive definite as the Hadamard product of two semi-positive definite matrices. To see that the first factor is indeed semi-positive definite one calculates the Schur complement;

$$B - C\hat{B}^{-1}C \propto B\hat{B} - C^2$$
$$= 64m^2(2m-1)^2|x|^{4m-4}|\theta|^{4m-4} + 32m^2(m-1)(2m-1)|x|^{2m-4}|\theta|^{6m-4}$$
$$+ 32m^2(m-1)(2m-1)|x|^{2m-4}|\theta|^{4m-4} + 32m^2(m-1)(2m-1)|x|^{6m-4}|\theta|^{2m-4}$$
$$+ 16m^2(m-1)^2|x|^{4m-4}|\theta|^{4m-4} + 16m^2(m-1)^2|x|^{4m-4}|\theta|^{2m-4}$$
$$+ 32m^2(m-1)(2m-1)|x|^{4m-4}|\theta|^{2m-4} + 16m^2(m-1)^2|x|^{2m-4}|\theta|^{4m-4}$$
$$+ 16m^2(m-1)^2|x|^{2m-4}|\theta|^{2m-4} - 16m^4|x|^{4m-4}|\theta|^{4m-4} \geq 0.$$

Likewise one argues that the rightmost matrix is also semi-positive definite, thus the Hessian $H(U(\theta, x))$ is positive definite and U strongly convex. $\square$

B.3. Mixed term - Toy Example

Remark 7. *Consider the potential* $U(\theta, x) = x\theta + |x|^4 + (|x|^2 + 1)(|\theta|^2 + 1) + |\theta|^4$, *it satisfies* A2, *i.e. it is strongly convex with* $\mu = 1$.

Proof. We decompose the Hessian of $U(\theta, x)$ as follows

$$H(U(\theta, x))$$
$$= \begin{bmatrix} (4|x|^2 + 2|\theta|^2)I_d & 0 \\ 0 & (4|\theta|^2 + 2|x|^2)I_d \end{bmatrix} + \begin{bmatrix} 2I_d & I_d \\ I_d & 2I_d \end{bmatrix} + \begin{bmatrix} 8J_d & 4J_d \\ 4J_d & 8J_d \end{bmatrix} \odot \begin{bmatrix} x \\ \theta \end{bmatrix} \begin{bmatrix} x & \theta \end{bmatrix},$$

One observes that the leftmost matrix is semi-positive definite due to its non-negative diagonal structure, while the rightmost matrix is also semi-positive definite as the product of a positive and semi-positive definite matrices. The intermediate term is positive definite and its smallest eigenvalue grants us $\mu = 1$ for the convexity parameter of $U(\theta, x)$. $\square$